Commun. Math. Phys. 349, 947–990 (2017)
Digital Object Identifier (DOI) 10.1007/s00220-016-2805-6

Communications in
Mathematical
Physics

Local Law of Addition of Random Matrices on Optimal Scale

Zhigang Bao[1,2], **László Erdős**[2], **Kevin Schnelli**[2,3]

[1] Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. E-mail: mazgbao@ust.hk
[2] IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria. E-mail: lerdos@ist.ac.at
[3] Department of Mathematics, KTH Royal Institute of Technology, Lindstedtsvägen 25, 100 44 Stockholm Sweden. E-mail: schnelli@kth.se

Received: 8 December 2015 / Accepted: 19 October 2016
Published online: 18 November 2016 – © The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract: The eigenvalue distribution of the sum of two large Hermitian matrices, when one of them is conjugated by a Haar distributed unitary matrix, is asymptotically given by the free convolution of their spectral distributions. We prove that this convergence also holds locally in the bulk of the spectrum, down to the optimal scales larger than the eigenvalue spacing. The corresponding eigenvectors are fully delocalized. Similar results hold for the sum of two real symmetric matrices, when one is conjugated by Haar orthogonal matrix.

1. Introduction

The pioneering work [31] of Voiculescu connected free probability with random matrices, as one of the most prominent examples for a noncommutative probability space is the space of Hermitian $N \times N$ matrices. On one hand, the law of the sum of two free random variables with laws μ_α and μ_β is given by the free additive convolution $\mu_\alpha \boxplus \mu_\beta$. On the other hand, in the case of Hermitian matrices, the law can be identified with the distribution of the eigenvalues. Thus the free additive convolution computes the eigenvalue distribution of the sum of two free Hermitian matrices. However, freeness is characterized by an infinite collection of moment identities and cannot easily be verified in general. A fundamental direct mechanism to generate freeness is conjugation by random unitary matrices. More precisely, two large Hermitian random matrices are asymptotically free if the unitary transfer matrix between their eigenbases is Haar distributed. The most important example is when the spectra of the two matrices are deterministic and the unitary conjugation is the sole source of randomness. In other words, if $A = A^{(N)}$ and $B = B^{(N)}$ are two sequences of deterministic $N \times N$ Hermitian matrices and U is a Haar distributed unitary, then A and UBU^* are asymptotically free in the large N limit and the asymptotic eigenvalue distribution of $A + UBU^*$ is given by the free additive convolution $\mu_A \boxplus \mu_B$ of the eigenvalue distributions of A and B.

Z. Bao, L. Erdős and K.Schnelli were supported by ERC Advanced Grant RANMAT No. 338804.

Since Voiculescu's first proof, several alternative approaches have been developed, see *e.g.* [11,16,29,30], but all of them were *global* in the sense that they describe the eigenvalue distribution in the weak limit, *i.e.* on the macroscopic scale, tested against N-independent test functions (to fix the scaling, we assume that $A^{(N)}$ and $B^{(N)}$ are uniformly bounded).

The study of a *local law*, *i.e.* identification of the eigenvalue distribution of $A+UBU^*$ with the free additive convolution below the macroscopic scale, was initiated by Kargin. First, he reached the scale $(\log N)^{-1/2}$ in [25] by using the Gromov–Milman concentration inequality for the Haar measure (a weaker concentration result was obtained earlier by Chatterjee [14]). Kargin later improved his result down to scale $N^{-1/7}$ in the bulk of the spectrum [26] by analyzing the stability of the subordination equations more efficiently. This result was valid only away from finitely many points in the bulk spectrum and no effective control was given on this exceptional set. Recently in [1], we reduced the minimal scale to $N^{-2/3}$ by establishing the optimal stability and by using a bootstrap procedure to successively localize the Gromov–Milman inequality from larger to smaller scales. Moreover, our result holds in the entire bulk spectrum. In fact, the key novelty in [1] was a new stability analysis in the entire bulk spectrum.

The main result of the current paper is the local law for $H = A+UBU^*$ down to the scale $N^{-1+\gamma}$, for any $\gamma > 0$. Note that the typical eigenvalue spacing is of order N^{-1}, a scale where the eigenvalue density fluctuates and no local law holds. Thus our result holds down to the optimal scale.

There are several motivations to establish such refinements of the macroscopic limit laws. First, such bounds are used as a priori estimates in the proofs of Wigner–Dyson–Mehta type universality results on local spectral statistics; see *e.g.* [12,20,21,27] and references therein. Second, control on the diagonal resolvent matrix elements for some $\eta = \operatorname{Im} z$ implies that the eigenvectors are delocalized on scale η^{-1}; the optimal scale for η yields complete delocalization of the eigenvectors. Third, the local law is ultimately related to an effective speed of convergence in Voiculescu's theorem on the global scale [1,26].

The basic idea of the proof is a continuity argument in the imaginary part $\eta = \operatorname{Im} z$ of the spectral parameter $z \in \mathbb{C}^+$ in the resolvent $G(z) = (H-z)^{-1}$. This method for the matrix elements of $G(z)$ was first introduced in [19] in the context of Wigner matrices. It requires an initial step, an a priori control on $G(z)$ for large η, say $\eta = 1$. In the context of the current paper, the a priori bound is provided by Kargin's result [26]. Since $G(z)$ is continuous in z, this also provides a control on $G(z)$ for slightly smaller η. This weak control shows that the normalized trace of $G(z)$ (and in fact all diagonal elements G_{ii}) is in the stability regime of a self-consistent equation which identifies the limiting object. The main work is to estimate the error between the equations for $G(z)$ and its limit. Our analysis has three major ingredients.

First, we use a *partial randomness decomposition* of the Haar measure that enables us to take partial expectation of G_{ii} with respect to the ith column of U. Second, to compute this partial expectation, we establish a new system of self-consistent equations involving *only two* auxiliary quantities. Keeping in mind, as a close analogy, that freeness involves checking infinitely many moment conditions for monomials of A, B and U, one may fear that an equation for G involves BG, whose equation involves BGB *etc.*, *i.e.* one would end up with an infinite system of equations. Surprisingly this is not the case and monitoring two appropriately chosen quantities in tandem is sufficient to close the system. Third, to connect the partial expectation of G_{ii} with the subordination functions from free probability, we rely on the optimal stability result for the subordination equations obtained in [1].

We stress that exploiting concentration only for the partial randomness surpasses the more general but less flexible Gromov–Milman technique. The main point is that we use concentration for each G_{ii} separately, exploiting the randomness of a single column (namely the ith one) of the Haar unitary U. Since G_{ii} depends much stronger on this column than on the other ones, the partial expectation of G_{ii} with respect to the ith column is already essentially deterministic. The concentration around this partial expectation is more efficient since it uses only $O(N)$ random variables instead of all the $O(N^2)$ variables used in Gromov-Milman method.

One prominent application of our work concerns the *single ring theorem* of Guionnet, Krishnapur and Zeitouni [22] on the eigenvalue distribution of matrices of the form UTV, where T is a fixed positive definite matrix and U, V are independent Haar distributed. Via the hermitization technique, local laws for the addition of random matrices can be used to prove local versions of the single ring theorem. This approach was demonstrated recently by Benaych-Georges [8], who proved a local single ring theorem on scale $(\log N)^{-1/4}$ using Kargin's local law on scale $(\log N)^{-1/2}$. The local law on the optimal scale N^{-1} is one of the key ingredients to prove the local single ring theorem on the optimal scale. The local single ring theorem will be proved in our separate work [2].

1.1. Notation. The following definition for high-probability estimates is suited for our purposes, which was first used in [18].

Definition 1.1. Let

$$X = (X^{(N)}(v) : N \in \mathbb{N}, v \in \mathcal{V}^{(N)}), \quad Y = (Y^{(N)}(v) : N \in \mathbb{N}, v \in \mathcal{V}^{(N)}) \tag{1.1}$$

be two families of nonnegative random variables where $\mathcal{V}^{(N)}$ is a possibly N-dependent parameter set. We say that Y stochastically dominates X, uniformly in v, if for all (small) $\epsilon > 0$ and (large) $D > 0$,

$$\sup_{v \in \mathcal{V}^{(N)}} \mathbb{P}\left(X^{(N)}(v) > N^{\epsilon} Y^{(N)}(v) \right) \leq N^{-D}, \tag{1.2}$$

for sufficiently large $N \geq N_0(\epsilon, D)$. If Y stochastically dominates X, uniformly in v, we write $X \prec Y$.

We further rely on the following notation. We use the symbols $O(\cdot)$ and $o(\cdot)$ for the standard big-O and little-o notation. We use c and C to denote strictly positive constants that do not depend on N. Their values may change from line to line. For $a, b \geq 0$, we write $a \lesssim b$, $a \gtrsim b$ if there is $C \geq 1$ such that $a \leq Cb$, $a \geq C^{-1}b$ respectively.

We use bold font for vectors in $\mathbb{C}^N$ and denote the components as $\boldsymbol{v} = (v_1, \ldots, v_N) \in \mathbb{C}^N$. The canonical basis of $\mathbb{C}^N$ is denoted by $(\boldsymbol{e}_i)_{i=1}^N$. For $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N$, we write $\boldsymbol{v}^* \boldsymbol{w}$ for the scalar product $\sum_{i=1}^N \overline{v}_i w_i$. We denote by $\|\boldsymbol{v}\|_2$ the Euclidean norm and by $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$ the uniform norm of $\boldsymbol{v} \in \mathbb{C}^N$.

We denote by $M_N(\mathbb{C})$ the set of $N \times N$ matrices over $\mathbb{C}$. For $A \in M_N(\mathbb{C})$, we denote by $\|A\|$ its operator norm and by $\|A\|_2$ its Hilbert-Schmidt norm. The matrix entries of A are denoted by $A_{ij} = \boldsymbol{e}_i^* A \boldsymbol{e}_j$. We denote by $\mathrm{tr} A$ the normalized trace of A, *i.e.* $\mathrm{tr} A = \frac{1}{N} \sum_{i=1}^N A_{ii}$. For $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N$, the rank-one matrix $\boldsymbol{v}\boldsymbol{w}^*$ has elements $(\boldsymbol{v}\boldsymbol{w}^*)_{ij} = (v_i \overline{w}_j)$.

Let $\boldsymbol{g} = (g_1, \ldots, g_N)$ be a real or complex Gaussian vector. We write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ if $g_1, \ldots, g_N$ are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$ normal variables; and we write $\boldsymbol{g} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2 I_N)$ if $g_1, \ldots, g_N$ are i.i.d. $N_{\mathbb{C}}(0, \sigma^2)$ variables, where $g_i \sim N_{\mathbb{C}}(0, \sigma^2)$ means that $\operatorname{Re} g_i$ and $\operatorname{Im} g_i$ are independent $N(0, \frac{\sigma^2}{2})$ normal variables.

Finally, we use double brackets to denote index sets, *i.e.* ,

$$[\![n_1, n_2]\!] := [n_1, n_2] \cap \mathbb{Z}\,,$$

for $n_1, n_2 \in \mathbb{R}$.

2. Main Results

2.1. Free additive convolution. In this subsection, we recall the definition of the free additive convolution. This is a shortened version of Sect. 2.1 of [1] added for completeness.

Given a probability measure[1] μ on $\mathbb{R}$ its *Stieltjes transform*, m_μ, on the complex upper half-plane $\mathbb{C}^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ is defined by

$$m_\mu(z) := \int_{\mathbb{R}} \frac{\mathrm{d}\mu(x)}{x - z}\,, \quad z \in \mathbb{C}^+. \tag{2.1}$$

Note that $m_\mu : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function such that

$$\lim_{\eta \nearrow \infty} \mathrm{i}\eta\, m_\mu(\mathrm{i}\eta) = -1. \tag{2.2}$$

Conversely, if $m : \mathbb{C}^+ \to \mathbb{C}^+$ is an analytic function such that $\lim_{\eta \nearrow \infty} \mathrm{i}\eta\, m(\mathrm{i}\eta) = 1$, then m is the Stieltjes transform of a probability measure μ, *i.e.* $m(z) = m_\mu(z)$, for all $z \in \mathbb{C}^+$.

We denote by F_μ the *negative reciprocal Stieltjes transform* of μ, *i.e.*

$$F_\mu(z) := -\frac{1}{m_\mu(z)}\,, \quad z \in \mathbb{C}^+. \tag{2.3}$$

Observe that

$$\lim_{\eta \nearrow \infty} \frac{F_\mu(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1\,, \tag{2.4}$$

as follows from (2.2), and note that F_μ is analytic on $\mathbb{C}^+$ with nonnegative imaginary part.

The *free additive convolution* is the symmetric binary operation on probability measures on $\mathbb{R}$ characterized by the following result.

Proposition 2.1 (Theorem 4.1 in [6], Theorem 2.1 in [15]). *Given two probability measures, μ_1 and μ_2, on $\mathbb{R}$, there exist unique analytic functions, $\omega_1\,, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, such that,*

[1] All probability measures considered will be assumed to be Borel.

(i) *for all* $z \in \mathbb{C}^+$, $\operatorname{Im}\omega_1(z),\ \operatorname{Im}\omega_2(z) \geq \operatorname{Im} z$, *and*

$$\lim_{\eta\nearrow\infty} \frac{\omega_1(\mathrm{i}\eta)}{\mathrm{i}\eta} = \lim_{\eta\nearrow\infty} \frac{\omega_2(\mathrm{i}\eta)}{\mathrm{i}\eta} = 1\,; \tag{2.5}$$

(ii) *for all* $z \in \mathbb{C}^+$,

$$F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z))\,, \quad \omega_1(z) + \omega_2(z) - z = F_{\mu_1}(\omega_2(z)). \tag{2.6}$$

It follows from (2.5) that the analytic function $F : \mathbb{C}^+ \to \mathbb{C}^+$ defined by

$$F(z) := F_{\mu_1}(\omega_2(z)) = F_{\mu_2}(\omega_1(z)), \tag{2.7}$$

satisfies the analogue of (2.4). Thus F is the negative reciprocal Stieltjes transform of a probability measure μ, called the free additive convolution of μ_1 and μ_2, usually denoted by $\mu \equiv \mu_1 \boxplus \mu_2$. The functions ω_1 and ω_2 of Proposition 2.1 are called *subordination functions* and m is said to be subordinated to m_{μ_1}, respectively to m_{μ_2}. Moreover, observe that ω_1 and ω_2 are analytic functions on $\mathbb{C}^+$ with nonnegative imaginary parts. Hence they admit the Nevanlinna representations

$$\omega_j(z) = a_{\omega_j} + z + \int_{\mathbb{R}} \frac{1+zx}{x-z}\,\mathrm{d}\varrho_{\omega_j}(x)\,, \quad j = 1, 2\,, \qquad z \in \mathbb{C}^+\,, \tag{2.8}$$

where $a_{\omega_j} \in \mathbb{R}$ and ϱ_{ω_j} are finite Borel measures on $\mathbb{R}$. For further details and historical remarks on the free additive convolution we refer to, *e.g.* [23,32].

Choosing μ_1 as a single point mass at $b \in \mathbb{R}$ and μ_2 arbitrary, it is straightforward to check that $\mu_1 \boxplus \mu_2$ is μ_2 shifted by b. We exclude this uninteresting case by assuming hereafter that μ_1 and μ_2 are both supported at more than one point. For general μ_1 and μ_2, the atoms of $\mu_1 \boxplus \mu_2$ are identified as follows. A point $c \in \mathbb{R}$ is an atom of $\mu_1 \boxplus \mu_2$, if and only if there exist $a, b \in \mathbb{R}$ such that $c = a + b$ and $\mu_1(\{a\}) + \mu_2(\{b\}) > 1$; see [Theorem 7.4, [10]]. Properties of the continuous part of $\mu_1 \boxplus \mu_2$ may be inferred from the boundary behavior of the functions $F_{\mu_1\boxplus\mu_2}$, ω_1 and ω_2. For simplicity, we restrict the discussion to compactly supported probability measures in the following.

Proposition 2.2 (Theorem 2.3 in [3], Theorem 3.3 in [4]). *Let μ_1 and μ_2 be compactly supported probability measures on $\mathbb{R}$ none of them being a single point mass. Then the functions $F_{\mu_1\boxplus\mu_2}$, ω_1, $\omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$ extend continuously to $\mathbb{R}$.*

Belinschi further showed in Theorem 4.1 in [4] that the singular continuous part of $\mu_1 \boxplus \mu_2$ is always zero and that the absolutely continuous part, $(\mu_1 \boxplus \mu_2)^{\mathrm{ac}}$, of $\mu_1 \boxplus \mu_2$ is always nonzero. We denote the density function of $(\mu_1 \boxplus \mu_2)^{\mathrm{ac}}$ by $f_{\mu_1\boxplus\mu_2}$.

We are now all set to introduce our notion of *regular bulk*, $\mathcal{B}_{\mu_1\boxplus\mu_2}$, of $\mu_1 \boxplus \mu_2$. Informally, we let $\mathcal{B}_{\mu_1\boxplus\mu_2}$ be the open set on which $\mu_1 \boxplus \mu_2$ has a continuous density that is strictly positive and bounded from above. For a formal definition we first introduce the set

$$\mathcal{U}_{\mu_1\boxplus\mu_2} := \operatorname{int}\left\{ \operatorname{supp}(\mu_1 \boxplus \mu_2)^{\mathrm{ac}} \setminus \{x \in \mathbb{R} : \lim_{\eta\searrow 0} F_{\mu_1\boxplus\mu_2}(x + \mathrm{i}\eta) = 0\} \right\}. \tag{2.9}$$

Note that $\mathcal{U}_{\mu_1\boxplus\mu_2}$ does not contain any atoms of $\mu_1 \boxplus \mu_2$. By the Luzin–Privalov theorem the set $\{x \in \mathbb{R} : \lim_{\eta\searrow 0} F_{\mu_1\boxplus\mu_2}(x + \mathrm{i}\eta) = 0\}$ has Lebesgue measure zero. In fact, a stronger statement applies for the case at hand. Belinschi [5] showed that if $x \in \mathbb{R}$ is such that $\lim_{\eta\searrow 0} F_{\mu_1\boxplus\mu_2}(x + \mathrm{i}\eta) = 0$, then it must be of the form $x = a + b$ with $\mu_1(\{a\}) + \mu_2(\{b\}) \geq 1$, $a, b \in \mathbb{R}$. Since there could only be finitely many such point x, the set $\mathcal{U}_{\mu_1\boxplus\mu_2}$ must contain an open non-empty interval.

Proposition 2.3 (Theorem 3.3 in [4]). *Let μ_1 and μ_2 be as above and fix any $x \in \mathcal{U}_{\mu_1 \boxplus \mu_2}$. Then $F_{\mu_1 \boxplus \mu_2}$, ω_1, $\omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$ extend analytically around x. Thus the density function $f_{\mu_1 \boxplus \mu_2}$ is real analytic in $\mathcal{U}_{\mu_1 \boxplus \mu_2}$ wherever positive.*

The regular bulk is obtained from $\mathcal{U}_{\mu_1 \boxplus \mu_2}$ by removing the zeros of $f_{\mu_1 \boxplus \mu_2}$ inside $\mathcal{U}_{\mu_1 \boxplus \mu_2}$.

Definition 2.4. The regular bulk of the measure $\mu_1 \boxplus \mu_2$ is the set

$$\mathcal{B}_{\mu_1 \boxplus \mu_2} := \mathcal{U}_{\mu_1 \boxplus \mu_2} \setminus \left\{ x \in \mathcal{U}_{\mu_1 \boxplus \mu_2} \,:\, f_{\mu_1 \boxplus \mu_2}(x) = 0 \right\}. \tag{2.10}$$

Note that $\mathcal{B}_{\mu_1 \boxplus \mu_2}$ is an open nonempty set on which $\mu_1 \boxplus \mu_2$ admits the density $f_{\mu_1 \boxplus \mu_2}$. The density is strictly positive and thus by Proposition 2.3 real analytic on $\mathcal{B}_{\mu_1 \boxplus \mu_2}$.

2.2. Definition of the model and assumptions. Let $A \equiv A^{(N)}$ and $B \equiv B^{(N)}$ be two sequences of deterministic real diagonal matrices in $M_N(\mathbb{C})$, whose empirical spectral distributions are denoted by μ_A and μ_B, respectively. More precisely,

$$\mu_A := \frac{1}{N} \sum_{i=1}^{N} \delta_{a_i}, \qquad \mu_B := \frac{1}{N} \sum_{i=1}^{N} \delta_{b_i}, \tag{2.11}$$

with $A = \operatorname{diag}(a_i)$, $B = \operatorname{diag}(b_i)$. For simplicity we omit the N-dependence of the matrices A and B from our notation. Throughout the paper, we assume

$$\|A\|, \|B\| \le C, \tag{2.12}$$

for some positive constant C uniform in N.

Proposition 2.1 asserts the existence of unique analytic functions ω_A and ω_B satisfying the analogue of (2.5) such that, for all $z \in \mathbb{C}^+$,

$$F_{\mu_A}(\omega_B(z)) = F_{\mu_B}(\omega_A(z)), \qquad \omega_A(z) + \omega_B(z) - z = F_{\mu_A}(\omega_B(z)). \tag{2.13}$$

We will assume that there are deterministic probability measures μ_α and μ_β on $\mathbb{R}$, neither of them being a single point mass, such that the empirical spectral distributions μ_A and μ_B converge weakly to μ_α and μ_β, as $N \to \infty$. More precisely, we assume that

$$\mathrm{d_L}(\mu_A, \mu_\alpha) + \mathrm{d_L}(\mu_B, \mu_\beta) \to 0, \tag{2.14}$$

as $N \to \infty$, where $\mathrm{d_L}$ denotes the Lévy distance. Proposition 2.1 asserts that there are unique analytic functions ω_α, ω_β satisfying the analogue of (2.5) such that, for all $z \in \mathbb{C}^+$,

$$F_{\mu_\alpha}(\omega_\beta(z)) = F_{\mu_\beta}(\omega_\alpha(z)), \qquad \omega_\alpha(z) + \omega_\beta(z) - z = F_{\mu_\alpha}(\omega_\beta(z)). \tag{2.15}$$

Proposition 4.13 of [9] states that $\mathrm{d_L}(\mu_A \boxplus \mu_B, \mu_\alpha \boxplus \mu_\beta) \le \mathrm{d_L}(\mu_A, \mu_\alpha) + \mathrm{d_L}(\mu_B, \mu_\beta)$, *i.e.* the free additive convolution is continuous with respect to weak convergence of measures.

Denote by $U(N)$ the unitary group of degree N. Let $U \in U(N)$ be distributed according to the Haar measure (in short U is a *Haar unitary*), and consider the random matrix

$$H \equiv H^{(N)} := A + UBU^*. \tag{2.16}$$

Our results also hold for the real case when U is Haar distributed on the orthogonal group, $O(N)$, of degree N. Throughout the main part of the paper the discussion will focus on the unitary case while the orthogonal case is addressed in Appendix A.

We introduce the *Green function*, G_H, of H and its normalized trace, m_H, by

$$G_H(z) := \frac{1}{H - z}, \qquad m_H(z) := \operatorname{tr} G_H(z), \qquad z \in \mathbb{C}^+. \tag{2.17}$$

For simplicity, we frequently use the notation $G(z)$ instead of $G_H(z)$ and we write $G_{ij}(z) \equiv (G_H)_{ij}(z)$ for the (i, j)th matrix element of $G(z)$.

2.3. Main results. For $a, b \geq 0$, $b \geq a$, and $\mathcal{I} \subset \mathbb{R}$, let

$$\mathcal{S}_{\mathcal{I}}(a, b) := \{z = E + \mathrm{i}\eta \in \mathbb{C}^+ : E \in \mathcal{I}, a \leq \eta \leq b\}, \tag{2.18}$$

In addition, for brevity, we set, for any given $\gamma > 0$,

$$\eta_{\mathrm{m}} \equiv \eta_{\mathrm{m}}(\gamma) := N^{-1+\gamma}. \tag{2.19}$$

The main results of this paper are as follows.

Theorem 2.5. *Let μ_α and μ_β be two compactly supported probability measures on $\mathbb{R}$, and assume that neither is supported at a single point and that at least one of them is supported at more than two points. Assume that the sequences of matrices A and B in (2.16) are such that their empirical eigenvalue distributions μ_A and μ_B satisfy (2.14). Let $\mathcal{I} \subset \mathcal{B}_{\mu_\alpha \boxplus \mu_\beta}$ be a nonempty compact interval.*

Then, for any fixed $\gamma > 0$, the estimates

$$\max_{1 \leq i \leq N} \left| G_{ii}(z) - \frac{1}{a_i - \omega_B(z)} \right| \prec \frac{1}{\sqrt{N\eta}}, \tag{2.20}$$

$$\max_{i \neq j} \left| G_{ij}(z) \right| \prec \frac{1}{\sqrt{N\eta}} \tag{2.21}$$

and

$$\left| m_H(z) - m_{\mu_A \boxplus \mu_B}(z) \right| \prec \frac{1}{\sqrt{N\eta}} \tag{2.22}$$

hold uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ (see (2.18)), where $\eta \equiv \operatorname{Im} z$ and $\eta_{\mathrm{m}} \equiv \eta_{\mathrm{m}}(\gamma)$ is given in (2.19).

Remark 2.1. The assumption that neither of μ_α and μ_β is a point mass, ensures that the free additive convolution is not a simple translate. The additional assumption that at least one of them is supported at more than two points is made for brevity of the exposition here. In Appendix B, we present the corresponding result for the special case when μ_α and μ_β are both convex combinations of two point masses.

Remark 2.2. We recall from Lemma 5.1 and Theorem 2.7 of [1] that, under the conditions of Theorem 2.5, there is a finite constant C such that

$$\max_{z \in \mathcal{S}_{\mathcal{I}}(0,1)} \max \left\{ \left| \omega_A(z) - \omega_\alpha(z) \right|, \left| \omega_B(z) - \omega_\beta(z) \right|, \left| m_{\mu_A \boxplus \mu_B}(z) - m_{\mu_\alpha \boxplus \mu_\beta}(z) \right| \right\}$$
$$\leq C \left(\mathrm{d_L}(\mu_A, \mu_\alpha) + \mathrm{d_L}(\mu_B, \mu_\beta) \right), \tag{2.23}$$

i.e. the Lévy distances of the empirical eigenvalue distributions of A and B from their limiting distributions control uniformly the deviations of the corresponding subordination

functions and Stieltjes transforms. Note moreover that $\max_{z\in\mathcal{S}_{\mathcal{I}}(0,1)}|m_{\mu_\alpha\boxplus\mu_\beta}(z)|<\infty$ by compactness of $\mathcal{I}$ and analyticity of $m_{\mu_1\boxplus\mu_2}$. Thus the Stieltjes-Perron inversion formula directly implies that $(\mu_A\boxplus\mu_B)^{\mathrm{ac}}$ has a density, $f_{\mu_A\boxplus\mu_B}$, inside $\mathcal{I}$ and that

$$\max_{x\in\mathcal{I}}|f_{\mu_A\boxplus\mu_B}(x)-f_{\mu_\alpha\boxplus\mu_\beta}(x)|\le C\left(\mathrm{d_L}(\mu_A,\mu_\alpha)+\mathrm{d_L}(\mu_B,\mu_\beta)\right), \tag{2.24}$$

for N sufficiently large. In particular, since $\mathcal{I}\subset\mathcal{B}_{\mu_\alpha\boxplus\mu_\beta}$, we have $\mathcal{I}\subset\mathcal{B}_{\mu_A\boxplus\mu_B}$, for N sufficiently large, *i.e.* we use, for large N, $\mu_\alpha\boxplus\mu_\beta$ to locate (an interval of) the regular bulk of $\mu_A\boxplus\mu_B$. Combining (2.23) and (2.20), we get

$$\max_{1\le i\le N}\left|G_{ii}(z)-\frac{1}{a_i-\omega_\beta(z)}\right|\prec\frac{1}{\sqrt{N\eta}}+\mathrm{d_L}(\mu_A,\mu_\alpha)+\mathrm{d_L}(\mu_B,\mu_\beta), \tag{2.25}$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}},1)$, where $\eta=\operatorname{Im}z$. Hereafter, we use the abbreviation m_A,m_B, m_α,m_β for $m_{\mu_A},m_{\mu_B},m_{\mu_\alpha},m_{\mu_\beta}$ respectively. Averaging over the index i, we get the corresponding statement for $|m_H(z)-m_A(\omega_\beta(z))|$ with the same error bound. Further, a lower bound on $\operatorname{Im}\omega_\beta(z)$ (*c.f.* (3.15)) implies $|m_A(\omega_\beta(z))-m_\alpha(\omega_\beta(z))|\lesssim\mathrm{d_L}(\mu_A,\mu_\alpha)$. Hence, using $m_\alpha(\omega_\beta(z))=m_{\mu_\alpha\boxplus\mu_\beta}(z)$, we observe that $|m_H(z)-m_{\mu_A\boxplus\mu_B}(z)|$ is bounded by the right side of (2.25), too.

Remark 2.3. Note that assumption (2.14) does not exclude that the matrix H has outliers in the large N limit. In fact, the model $H=A+UBU^*$ shows a rich phenomenology when, say, A has a finite number of large spikes; we refer to the recent works in [7,13,26].

Let $\lambda_1,\dots,\lambda_N$ be the eigenvalues of H, and $\boldsymbol{u}_1,\dots,\boldsymbol{u}_N$ be the corresponding ℓ^2-normalized eigenvectors. The following result shows complete delocalization of the bulk eigenvectors.

Theorem 2.6 (Delocalization of eigenvectors). *Under the assumptions of Theorem 2.5 the following holds. Let $\mathcal{I}\subset\mathcal{B}_{\mu_\alpha\boxplus\mu_\beta}$ be a compact nonempty interval. Then*

$$\max_{i:\lambda_i\in\mathcal{I}}\|\boldsymbol{u}_i\|_\infty\prec\frac{1}{\sqrt{N}}. \tag{2.26}$$

2.4. Strategy of proof. In this subsection, we informally outline the strategy of our proofs. Throughout the paper, without loss of generality, we assume

$$\operatorname{tr}A=\operatorname{tr}B=0. \tag{2.27}$$

For brevity, we use the shorthand $m_\boxplus\equiv m_{\mu_A\boxplus\mu_B}$ for the Stieltjes transform of $\mu_A\boxplus\mu_B$.

We consider first the unitary setting. Let

$$H:=A+UBU^*,\qquad\mathcal{H}:=U^*AU+B, \tag{2.28}$$

and denote their Green functions by

$$G(z)=(H-z)^{-1},\qquad\mathcal{G}(z)=(\mathcal{H}-z)^{-1},\quad z\in\mathbb{C}^+. \tag{2.29}$$

We write $z=E+\mathrm{i}\eta\in\mathbb{C}^+$, $E\in\mathbb{R}$ and $\eta>0$, for the spectral parameter. In the sequel we often omit $z\in\mathbb{C}^+$ from the notation when no confusion can arise. Recalling (2.17), we have

$$m_H(z)=\operatorname{tr}G(z)=\operatorname{tr}\mathcal{G}(z),\quad z\in\mathbb{C}^+.$$

For brevity, we set

$$\widetilde{A} := U^* A U\,, \qquad \widetilde{B} := U B U^*.$$

The following functions will play a key role in our proof.

Definition 2.7 (Approximate subordination functions).

$$\omega_A^c(z) := z - \frac{\operatorname{tr} \widetilde{A}\mathcal{G}(z)}{m_H(z)}\,, \qquad \omega_B^c(z) := z - \frac{\operatorname{tr} \widetilde{B}G(z)}{m_H(z)}\,, \qquad z \in \mathbb{C}^+. \tag{2.30}$$

Notice that the role of A and B are not symmetric in these notations. By cyclicity of the trace, we may write

$$\omega_A^c(z) = z - \frac{\operatorname{tr} AG(z)}{m_H(z)}\,, \qquad z \in \mathbb{C}^+. \tag{2.31}$$

We remark that the approximate subordination functions defined above are slightly different from the candidate subordination functions used in [26,29] which were later used in [1].

The functions $\omega_A^c(z)$ and $\omega_B^c(z)$ turn out to be good approximations to the subordination functions $\omega_A(z)$ and $\omega_B(z)$ of (2.13). A direct consequence of the definition in (2.30) is that

$$\frac{1}{m_H(z)} = z - \omega_A^c(z) - \omega_B^c(z)\,, \qquad z \in \mathbb{C}^+. \tag{2.32}$$

Having set the notation, our main task is to show that

$$G_{ii}(z) = \frac{1}{a_i - \omega_B^c(z)} + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big)\,, \qquad z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)\,, \tag{2.33}$$

where we focus, for simplicity, on the diagonal Green function entries only.

We first heuristically explain how (2.33) leads to our main result in (2.20). A key input is the local stability of the system (2.13) established in [1]; see Subsection 3.3 for a summary. Averaging over the index i in (2.33), we get

$$m_H(z) = m_A(\omega_B^c(z)) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big)\,, \tag{2.34}$$

with the shorthand notation $m_A(\,\cdot\,) \equiv m_{\mu_A}(\,\cdot\,)$. Replacing H by $\mathcal{H}$, we analogously get

$$m_H(z) = m_B(\omega_A^c(z)) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big)\,, \tag{2.35}$$

according to (2.31). Substituting (2.32) into (2.34) and (2.35) we obtain the system

$$F_{\mu_A}(\omega_B^c(z)) = \omega_A^c(z) + \omega_B^c(z) - z + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big),$$
$$F_{\mu_B}(\omega_A^c(z)) = \omega_A^c(z) + \omega_B^c(z) - z + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big),$$

which is a perturbation of (2.13). Using the local stability of the system (2.13), we obtain

$$\left|\omega_A^c(z) - \omega_A(z)\right| \prec \frac{1}{\sqrt{N\eta}}, \qquad \left|\omega_B^c(z) - \omega_B(z)\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{2.36}$$

Plugging the first estimate back into (2.33) we get (2.20). The full proof of this step is accomplished in Sect. 7.

We next return to (2.33). Its proof relies on the following decomposition of the Haar measure on the unitary group given, *e.g.* in [17,28]. For any fixed $i \in [\![1, N]\!]$, any Haar unitary U can be written as

$$U = -\mathrm{e}^{\mathrm{i}\theta_i} R_i \, U^{\langle i \rangle}. \tag{2.37}$$

Here R_i is the *Householder reflection* (up to a sign) sending the vector $\boldsymbol{e}_i$ to $\boldsymbol{v}_i$, where $\boldsymbol{v}_i \in \mathbb{C}^N$ is a random vector distributed uniformly on the complex unit $(N-1)$-sphere, and $\theta_i \in [0, 2\pi)$ is the argument of the ith coordinate of $\boldsymbol{v}_i$. The unitary matrix $U^{\langle i \rangle}$ has $\boldsymbol{e}_i$ as its ith column and its (i,i)-matrix minor (obtained by removing the ith column and ith row) is Haar distributed on $U(N-1)$; see Sect. 4 for more detail.

The gist of the decomposition in (2.37) is that the Householder reflection R_i and the unitary $U^{\langle i \rangle}$ are independent, for each fixed $i \in [\![1, N]\!]$. Hence, the decomposition in (2.37) allows one to split off the partial randomness of the vector $\boldsymbol{v}_i$ from U.

The proof of (2.33) is divided into two parts:

(i) Concentration of G_{ii} around the partial expectation $\mathbb{E}_{\boldsymbol{v}_i}[G_{ii}]$, *i.e.*

$$\left|G_{ii}(z) - \mathbb{E}_{\boldsymbol{v}_i}[G_{ii}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}.$$

(ii) Computation of the partial expectation $\mathbb{E}_{\boldsymbol{v}_i}[G_{ii}]$, *i.e.*

$$\left|\mathbb{E}_{\boldsymbol{v}_i}\big[G_{ii}(z)\big] - \big(a_i - \omega_B^c(z)\big)^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}.$$

To prove part (i), we resolve dependences by expansion and use concentration estimates for the vector $\boldsymbol{v}_i$. This part is accomplished in Sect. 5.

Part (ii) is carried out in Sect. 6. We start from the Green function identity

$$(a_i - z)G_{ii}(z) = -(\widetilde{B}G(z))_{ii} + 1. \tag{2.38}$$

Taking the $\mathbb{E}_{\boldsymbol{v}_i}$ expectation of (2.38) and recalling the definition of the approximate subordination function $\omega_B^c(z)$ in (2.30), it suffices to show that

$$\mathbb{E}_{\boldsymbol{v}_i}\big[(\widetilde{B}G)_{ii}\big] = \frac{\operatorname{tr}\widetilde{B}G}{\operatorname{tr}G} G_{ii} + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big),$$

to prove (2.33). Denoting $\widetilde{B}^{\langle i \rangle} := U^{\langle i \rangle} B (U^{\langle i \rangle})^*$ and setting, for $z \in \mathbb{C}^+$,

$$S_i^\sharp(z) := \mathrm{e}^{\mathrm{i}\theta_i} \boldsymbol{v}_i^* \widetilde{B}^{\langle i \rangle} G(z) \boldsymbol{e}_i, \quad T_i^\sharp(z) := \mathrm{e}^{\mathrm{i}\theta_i} \boldsymbol{v}_i^* G(z) \boldsymbol{e}_i,$$

we will prove that

$$\mathbb{E}_{\boldsymbol{v}_i}\big[(\widetilde{B}G(z))_{ii}\big] = -\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp(z)\big] + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N}}\Big). \tag{2.39}$$

Hence, it suffices to determine $\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big]$ instead. Approximating $\mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i$ by a Gaussian vector and using integration by parts for Gaussian random variables, we get the pair of equations

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] &= \operatorname{tr}\big(\widetilde{B}G\big)\big(\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] - b_i\,\mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big]\big)\\ &\quad + \operatorname{tr}\big(\widetilde{B}G\widetilde{B}\big)\big(G_{ii} + \mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big]\big) + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N\eta}}\Big),\\ \mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big] &= \operatorname{tr}\big(G\big)\big(\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] - b_i\,\mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big]\big) + \operatorname{tr}\big(\widetilde{B}G\big)\big(G_{ii} + \mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big]\big) + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N\eta}}\Big),\end{aligned}$$

where we dropped the z-argument for the sake of brevity; see (6.23) and (6.24) for precise statements with, for technical reasons, slightly modified $S_i^\sharp$ and $T_i^\sharp$. Solving the two equations above for $\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big]$ we find

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] = -\frac{\operatorname{tr}\big(\widetilde{B}G\big)}{\operatorname{tr}G}G_{ii} + \bigg[\frac{\operatorname{tr}\big(\widetilde{B}G\big) - \big(\operatorname{tr}\big(\widetilde{B}G\big)\big)^2}{\operatorname{tr}G} + \operatorname{tr}\big(\widetilde{B}G\widetilde{B}\big)\bigg]\big(G_{ii} + \mathbb{E}_{\boldsymbol{v}_i}\big[T_i^\sharp\big]\big)\\ + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N\eta}}\Big).\end{aligned} \tag{2.40}$$

Returning to (2.39), we also obtain, using concentration estimates for $(\widetilde{B}G)_{ii}$ (which follow from the concentration estimates of G_{ii} established in part (i) and (2.38)), that

$$\Big|\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] + \operatorname{tr}\widetilde{B}G\Big| \prec \frac{1}{\sqrt{N\eta}}. \tag{2.41}$$

Thus, averaging (2.40) over the index i and comparing with (2.41), we conclude that

$$\Big|\frac{\operatorname{tr}\big(\widetilde{B}G\big) - \big(\operatorname{tr}\big(\widetilde{B}G\big)\big)^2}{\operatorname{tr}G} + \operatorname{tr}\big(\widetilde{B}G\widetilde{B}\big)\Big| \prec \frac{1}{\sqrt{N\eta}}.$$

Plugging this last estimate back into (2.40), we eventually find that

$$\mathbb{E}_{\boldsymbol{v}_i}\big[S_i^\sharp\big] = -\frac{\operatorname{tr}\big(\widetilde{B}G\big)}{\operatorname{tr}G}G_{ii} + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N\eta}}\Big),$$

which together with (2.39) and (2.38) gives us part (ii). This completes the sketch of the proof for the unitary case. The proof of the orthogonal case is similar. The necessary modifications are given in Appendix A.

3. Preliminaries

In this section, we first collect some basic tools used later on and then summarize results of [1]. In particular, we discuss, under the assumptions of Theorem 2.5, stability properties of the system (2.13) and state essential properties of the subordination functions ω_A and ω_B.

3.1. Stochastic domination and large deviation properties. Recall the definition of stochastic domination in Definition 1.1. The relation $\prec$ is a partial ordering: it is transitive and it satisfies the arithmetic rules of an order relation, *e.g.*, if $X_1 \prec Y_1$ and $X_2 \prec Y_2$ then $X_1+X_2 \prec Y_1+Y_2$ and $X_1X_2 \prec Y_1Y_2$. Further assume that $\Phi(v) \geq N^{-C}$ is deterministic and that $Y(v)$ is a nonnegative random variable satisfying $\mathbb{E}[Y(v)]^2 \leq N^{C'}$ for all v. Then $Y(v) \prec \Phi(v)$, uniformly in v, implies $\mathbb{E}[Y(v)] \prec \Phi(v)$, uniformly in v.

Gaussian vectors have well-known large deviation properties. We will use them in the following form whose proof is standard.

Lemma 3.1. *Let $X = (x_{ij}) \in M_N(\mathbb{C})$ be a deterministic matrix and let $\boldsymbol{y} = (y_i) \in \mathbb{C}^N$ be a deterministic complex vector. For a Gaussian random vector $\boldsymbol{g} = (g_1, \ldots, g_N) \in \mathcal{N}_{\mathbb{R}}(0, \sigma^2 I_N)$ or $\mathcal{N}_{\mathbb{C}}(0, \sigma^2 I_N)$, we have*

$$|\boldsymbol{y}^* \boldsymbol{g}| \prec \sigma \|\boldsymbol{y}\|_2 \,, \qquad |\boldsymbol{g}^* X \boldsymbol{g} - \sigma^2 N \operatorname{tr} X| \prec \sigma^2 \|X\|_2. \tag{3.1}$$

3.2. Rank-one perturbation formula. At various places, we use the following fundamental perturbation formula: for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{C}^N$ and an invertible $D \in M_N(\mathbb{C})$, we have

$$\left(D + \boldsymbol{\alpha}\boldsymbol{\beta}^*\right)^{-1} = D^{-1} - \frac{D^{-1}\boldsymbol{\alpha}\boldsymbol{\beta}^* D^{-1}}{1 + \boldsymbol{\beta}^* D^{-1} \boldsymbol{\alpha}} \,, \tag{3.2}$$

as can be checked readily. A standard application of (3.2) is recorded in the following lemma.

Lemma 3.2. *Let $D \in M_N(\mathbb{C})$ be Hermitian and let $Q \in M_N(\mathbb{C})$ be arbitrary. Then, for any finite-rank Hermitian matrix $R \in M_N(\mathbb{C})$, we have*

$$\left| \operatorname{tr}\left(Q\big(D + R - z\big)^{-1} \right) - \operatorname{tr}\left(Q(D - z)^{-1} \right) \right| \leq \frac{\operatorname{rank}(R) \|Q\|}{N\eta} \,, \qquad z = E + \mathrm{i}\eta \in \mathbb{C}^+. \tag{3.3}$$

Proof. Let $z \in \mathbb{C}^+$ and $\boldsymbol{\alpha} \in \mathbb{C}^N$. Then from (3.2) we have

$$\operatorname{tr}\left(Q\big(D \pm \boldsymbol{\alpha}\boldsymbol{\alpha}^* - z\big)^{-1} \right) - \operatorname{tr}\left(Q(D - z)^{-1} \right) = \mp \frac{1}{N} \frac{\boldsymbol{\alpha}^* (D - z)^{-1} Q (D - z)^{-1} \boldsymbol{\alpha}}{1 \pm \boldsymbol{\alpha}^* (D - z)^{-1} \boldsymbol{\alpha}}. \tag{3.4}$$

We can thus estimate

$$\begin{aligned} \left| \operatorname{tr}\left(Q\big(D \pm \boldsymbol{\alpha}\boldsymbol{\alpha}^* - z\big)^{-1} \right) - \operatorname{tr}\left(Q(D - z)^{-1} \right) \right| &\leq \frac{\|Q\|}{N} \frac{\|(D - z)^{-1} \boldsymbol{\alpha}\|_2^2}{\left| 1 \pm \boldsymbol{\alpha}^* (D - z)^{-1} \boldsymbol{\alpha} \right|} \\ &= \frac{\|Q\|}{N\eta} \frac{\boldsymbol{\alpha}^* \operatorname{Im}\,(D - z)^{-1} \boldsymbol{\alpha}}{\left| 1 \pm \boldsymbol{\alpha}^* (D - z)^{-1} \boldsymbol{\alpha} \right|} \\ &\leq \frac{\|Q\|}{N\eta}. \end{aligned} \tag{3.5}$$

Since $R = R^* \in M_N(\mathbb{C})$ has finite rank, we can write R as a finite sum of rank-one Hermitian matrices of the form $\pm \boldsymbol{\alpha}\boldsymbol{\alpha}^*$. Thus iterating (3.5) we get (3.3). □

3.3. Local stability of the system (2.13). We first consider (2.6) in a general setting: For generic probability measures μ_1, μ_2, let $\Phi_{\mu_1,\mu_2} : (\mathbb{C}^+)^3 \to \mathbb{C}^2$ be given by

$$\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) := \begin{pmatrix} F_{\mu_1}(\omega_2) - \omega_1 - \omega_2 + z \\ F_{\mu_2}(\omega_1) - \omega_1 - \omega_2 + z \end{pmatrix}, \tag{3.6}$$

where F_{μ_1}, F_{μ_2} are the negative reciprocal Stieltjes transforms of μ_1, μ_2; see (2.3). Considering μ_1, μ_2 as fixed, the equation

$$\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) = 0\,, \tag{3.7}$$

is equivalent to (2.6) and, by Proposition 2.1, there are unique analytic functions $\omega_1, \omega_2 : \mathbb{C}^+ \to \mathbb{C}^+$, $z \mapsto \omega_1(z), \omega_2(z)$ satisfying (2.5) that solve (3.7) in terms of z. Choosing $\mu_1 = \mu_\alpha$, $\mu_2 = \mu_\beta$ equation (3.7) is equivalent to (2.15); choosing $\mu_1 = \mu_A$, $\mu_2 = \mu_B$ it is equivalent to (2.13). When no confusion can arise, we simply write Φ for $\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z)$.

We call the system (3.7) *linearly S-stable at* (ω_1, ω_2) if

$$\Gamma_{\mu_1,\mu_2}(\omega_1, \omega_2) := \left\| \begin{pmatrix} -1 & F'_{\mu_1}(\omega_2) - 1 \\ F'_{\mu_2}(\omega_1) - 1 & -1 \end{pmatrix}^{-1} \right\| \le S\,, \tag{3.8}$$

for some positive constant S. In particular, the partial Jacobian matrix of (3.6) given by

$$\mathrm{D}\Phi(\omega_1, \omega_2) := \left(\frac{\partial \Phi}{\partial \omega_1}(\omega_1, \omega_2, z)\,,\ \frac{\partial \Phi}{\partial \omega_2}(\omega_1, \omega_2, z) \right) = \begin{pmatrix} -1 & F'_{\mu_1}(\omega_2) - 1 \\ F'_{\mu_2}(\omega_1) - 1 & -1 \end{pmatrix},$$

has a bounded inverse at (ω_1, ω_2). Note that $\mathrm{D}\Phi(\omega_1, \omega_2)$ is independent of z. The implicit function theorem reveals that, if (3.7) is linearly S-stable at (ω_1, ω_2), then

$$\max_{z \in \mathcal{S}_{\mathcal{I}}(0,1)} |\omega_1'(z)| \le 2S\,, \qquad \max_{z \in \mathcal{S}_{\mathcal{I}}(0,1)} |\omega_2'(z)| \le 2S. \tag{3.9}$$

In particular, ω_1 and ω_2 are Lipschitz continuous with constant $2S$. A more detailed analysis yields the following local stability result of the system $\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) = 0$.

Lemma 3.3 (Proposition 4.1, [1]). *Fix $z_0 \in \mathbb{C}^+$. Assume that the functions $\widetilde{\omega}_1$, $\widetilde{\omega}_2$, $\widetilde{r}_1$, $\widetilde{r}_2 : \mathbb{C}^+ \to \mathbb{C}$ satisfy $\operatorname{Im} \widetilde{\omega}_1(z_0) > 0$, $\operatorname{Im} \widetilde{\omega}_2(z_0) > 0$ and*

$$\Phi_{\mu_1,\mu_2}(\widetilde{\omega}_1(z_0), \widetilde{\omega}_2(z_0), z_0) = \widetilde{r}(z_0)\,, \tag{3.10}$$

where $\widetilde{r}(z) := (\widetilde{r}_1(z), \widetilde{r}_2(z))^\top$. Assume moreover that there is $\delta \in [0, 1]$ such that

$$|\widetilde{\omega}_1(z_0) - \omega_1(z_0)| \le \delta\,, \qquad |\widetilde{\omega}_2(z_0) - \omega_2(z_0)| \le \delta\,, \tag{3.11}$$

where $\omega_1(z), \omega_2(z)$ solve the unperturbed system $\Phi_{\mu_1,\mu_2}(\omega_1, \omega_2, z) = 0$ with $\operatorname{Im} \omega_1(z) \ge \operatorname{Im} z$ and $\operatorname{Im} \omega_2(z) \ge z$, $z \in \mathbb{C}^+$. Assume that there is a constant S such that Φ is linearly S-stable at $(\omega_1(z_0), \omega_2(z_0))$, and assume in addition that there are strictly positive constants K and k with $k > \delta$ and with $k^2 > \delta K S$ such that

$$k \le \operatorname{Im} \omega_1(z_0) \le K\,, \qquad k \le \operatorname{Im} \omega_2(z_0) \le K. \tag{3.12}$$

Then

$$|\widetilde{\omega}_1(z_0) - \omega_1(z_0)| \le 2S \|\widetilde{r}(z_0)\|_2\,, \qquad |\widetilde{\omega}_2(z_0) - \omega_2(z_0)| \le 2S \|\widetilde{r}(z_0)\|_2. \tag{3.13}$$

In Sect. 7, we will apply Lemma 3.3 with the choices $\mu_1 = \mu_A$ and $\mu_2 = \mu_B$. We thus next show that the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$ is S-stable, for all $z \in \mathcal{S}_{\mathcal{I}}(0,1)$, and that (3.12) holds uniformly on $\mathcal{S}_{\mathcal{I}}(0,1)$; see (2.18) for the definition.

Lemma 3.4 (Lemma 5.1 and Corollary 5.2 of [1]). *Let μ_A, μ_B be the probability measures from (2.11) satisfying the assumptions of Theorem 2.5. Let ω_A, ω_B denote the associated subordination functions of (2.13). Let $\mathcal{I}$ be the interval in Theorem 2.5. Then for N sufficiently large, the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$ is S-stable with some positive constant S, uniformly on $\mathcal{S}_{\mathcal{I}}(0,1)$. Moreover, there exist two strictly positive constants K and k, such that for N sufficiently large, we have*

$$\max_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} |\omega_A(z)| \le K\,, \qquad \max_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} |\omega_B(z)| \le K\,, \tag{3.14}$$

$$\min_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} \operatorname{Im}\omega_A(z) \ge k\,, \qquad \min_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} \operatorname{Im}\omega_B(z) \ge k. \tag{3.15}$$

Remark 3.1. Under the assumptions of Lemma 3.4, the estimates in (3.15) can be extended as follows. There is $\tilde{k} > 0$ such that

$$\min_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} (\operatorname{Im}\omega_A(z) - \operatorname{Im} z) \ge \tilde{k}\,, \qquad \min_{z\in\mathcal{S}_{\mathcal{I}}(0,1)} (\operatorname{Im}\omega_B(z) - \operatorname{Im} z) \ge \tilde{k}. \tag{3.16}$$

This follows by combining (3.15) with the Nevanlinna representations in (2.8).

We conclude this section by mentioning that the general perturbation result in Lemma 3.3 combined with Lemma 3.4, can be used to prove (2.23). We refer to [1] for details.

4. Partial Randomness Decomposition

We use a decomposition of Haar measure on the unitary groups obtained in [17] (see also [28]): For a Haar distributed unitary matrix $U \equiv U_N$, there exist a random vector $\boldsymbol{v}_1 = (v_{11}, \dots, v_{1N})$, uniformly distributed on the complex unit $(N-1)$-sphere $\mathcal{S}_{\mathbb{C}}^{N-1} := \{\boldsymbol{x} \in \mathbb{C}^N : \boldsymbol{x}^*\boldsymbol{x} = 1\}$, and a Haar distributed unitary matrix $U^1 \equiv U^1_{N-1} \in U(N-1)$, which is independent of $\boldsymbol{v}_1$, such that one has the decomposition

$$U = -\mathrm{e}^{\mathrm{i}\theta_1}(I - \boldsymbol{r}_1\boldsymbol{r}_1^*)\begin{pmatrix} 1 & \\ & U^1 \end{pmatrix} =: -\mathrm{e}^{\mathrm{i}\theta_1} R_1 U^{\langle 1\rangle}\,,$$

where

$$\boldsymbol{r}_1 := \sqrt{2}\frac{\boldsymbol{e}_1 + \mathrm{e}^{-\mathrm{i}\theta_1}\boldsymbol{v}_1}{\|\boldsymbol{e}_1 + \mathrm{e}^{-\mathrm{i}\theta_1}\boldsymbol{v}_1\|_2}\,, \qquad R_1 := I - \boldsymbol{r}_1\boldsymbol{r}_1^*\,, \tag{4.1}$$

and where θ_1 is the argument of the first coordinate of the vector $\boldsymbol{v}_1$. More generally, for any $i \in [\![1, N]\!]$, there exists an independent pair $(\boldsymbol{v}_i, U^i)$, with $\boldsymbol{v}_i$ a uniformly distributed unit vector $\boldsymbol{v}_i$ and with $U^i \in U(N-1)$ a Haar unitary, such that one has the decomposition

$$U = -\mathrm{e}^{\mathrm{i}\theta_i} R_i U^{\langle i\rangle}\,, \qquad \boldsymbol{r}_i := \sqrt{2}\frac{\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|_2}\,, \qquad R_i := I - \boldsymbol{r}_i\boldsymbol{r}_i^*\,, \tag{4.2}$$

where $U^{\langle i\rangle}$ is the unitary matrix with $\boldsymbol{e}_i$ as its ith column and U^i as its (i,i)-matrix minor, and θ_i is the argument of the ith coordinate of $\boldsymbol{v}_i$. In addition, using the definition of R_i and $U^{\langle i\rangle}$, we note that $U\boldsymbol{e}_i = -\mathrm{e}^{\mathrm{i}\theta_i} R_i U^{\langle i\rangle}\boldsymbol{e}_i = \boldsymbol{v}_i$, *i.e.* $\boldsymbol{v}_i$ is the ith column of U.

With the above notation, we can write

$$H = A + R_i \widetilde{B}^{\langle i \rangle} R_i^* ,$$

for any $i \in [\![1, N]\!]$, where we introduced the shorthand notation

$$\widetilde{B}^{\langle i \rangle} := U^{\langle i \rangle} B \big(U^{\langle i \rangle}\big)^* . \tag{4.3}$$

We further define

$$H^{\langle i \rangle} := A + \widetilde{B}^{\langle i \rangle} , \qquad G^{\langle i \rangle}(z) := (H^{\langle i \rangle} - z)^{-1} , \qquad z \in \mathbb{C}^+ . \tag{4.4}$$

Note that $B^{\langle i \rangle}$, $H^{\langle i \rangle}$ and $G^{\langle i \rangle}$ are independent of $\boldsymbol{v}_i$.

It is well known that for a uniformly distributed unit vector $\boldsymbol{v}_i \in \mathbb{C}^N$, there exists a Gaussian vector $\widetilde{\boldsymbol{g}}_i = (\widetilde{g}_{i1}, \ldots, \widetilde{g}_{iN}) \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1} I)$ such that

$$\boldsymbol{v}_i = \frac{\widetilde{\boldsymbol{g}}_i}{\|\widetilde{\boldsymbol{g}}_i\|_2} . \tag{4.5}$$

By definition, θ_i is also the argument of $\widetilde{g}_{ii}$. Set

$$g_{ik} := \mathrm{e}^{-\mathrm{i}\theta_i} \widetilde{g}_{ik} , \qquad k \neq i , \tag{4.6}$$

and introduce an $N_{\mathbb{C}}(0, N^{-1})$ variable g_{ii} which is independent of the unitary matrix U and of $\widetilde{\boldsymbol{g}}_i$. Then, we denote $\boldsymbol{g}_i := (g_{i1}, \ldots, g_{iN})$ and note $\boldsymbol{g}_i \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1} I)$. In addition, by definition, we have

$$\mathrm{e}^{-\mathrm{i}\theta_i} \boldsymbol{v}_i - \boldsymbol{g}_i = \frac{|\widetilde{g}_{ii}| - g_{ii}}{\|\widetilde{\boldsymbol{g}}_i\|_2} \boldsymbol{e}_i + \Big(\frac{1}{\|\widetilde{\boldsymbol{g}}_i\|_2} - 1 \Big) \boldsymbol{g}_i .$$

In subsequent estimates for G_{ij}, it is convenient to approximate $\boldsymbol{r}_i$ by

$$\boldsymbol{w}_i := \boldsymbol{e}_i + \boldsymbol{g}_i \tag{4.7}$$

in the decomposition $U = -\mathrm{e}^{\mathrm{i}\theta_i} R_i U^{\langle i \rangle}$, without changing the randomness of $U^{\langle i \rangle}$. To estimate the precision of this approximation, we require more notation: Let

$$W_i = W_i^* := I - \boldsymbol{w}_i \boldsymbol{w}_i^* , \qquad \widetilde{B}^{(i)} = W_i \widetilde{B}^{\langle i \rangle} W_i . \tag{4.8}$$

Correspondingly, we denote

$$H^{(i)} := A + \widetilde{B}^{(i)} , \qquad G^{(i)}(z) := (H^{(i)} - z)^{-1} , \qquad z \in \mathbb{C}^+ . \tag{4.9}$$

The following lemma shows that $\boldsymbol{r}_i$ can be replaced by $\boldsymbol{w}_i$ in Green function entries at the expense of an error that is below the precision we are interested in.

Lemma 4.1. *Fix $z = E + \mathrm{i}\eta \in \mathbb{C}^+$ and choose indices $i, j, k \in [\![1, N]\!]$. Suppose that*

$$\begin{aligned} &\max \Big\{ |G_{kk}(z)|, |G_{ij}^{(i)}(z)| \Big\} \prec 1 , \\ &\max \Big\{ |\boldsymbol{g}_i^* G^{(i)}(z) \boldsymbol{e}_j|, |\boldsymbol{g}_i^* \widetilde{B}^{\langle i \rangle} G^{(i)}(z) \boldsymbol{e}_j| \Big\} \prec 1 , \end{aligned} \tag{4.10}$$

hold. Then

$$\big| G_{kj}(z) - G_{kj}^{(i)}(z) \big| \prec \frac{1}{\sqrt{N\eta}} \tag{4.11}$$

holds, too.

Proof of Lemma 4.1. Fix $i, j, k \in [\![1, N]\!]$. We first note that

$$\boldsymbol{r}_i = \boldsymbol{w}_i + \delta_{1i}\boldsymbol{e}_i + \delta_{2i}\boldsymbol{g}_i\,,$$

where

$$\begin{aligned}\delta_{1i} &:= \left(\frac{\sqrt{2}}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|_2} - 1\right) + \frac{\sqrt{2}}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|_2}\frac{|\widetilde{g}_{ii}| - g_{ii}}{\|\widetilde{\boldsymbol{g}}_i\|_2}\,,\\ \delta_{2i} &:= \frac{\sqrt{2}}{\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|_2}\frac{1}{\|\widetilde{\boldsymbol{g}}_i\|_2} - 1.\end{aligned} \tag{4.12}$$

By the concentration inequalities in Lemma 3.1, and $g_{ii}, \widetilde{g}_{ii} \sim N_{\mathbb{C}}(0, N^{-1})$, we see that

$$\begin{aligned}\|\widetilde{\boldsymbol{g}}_i\|_2 &= 1 + O_{\prec}(\frac{1}{\sqrt{N}}),\\ \|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\theta_i}\boldsymbol{v}_i\|_2 &= \left(2 + 2\frac{|\widetilde{g}_{ii}|}{\|\widetilde{\boldsymbol{g}}_i\|_2}\right)^{\frac{1}{2}} = \sqrt{2} + O_{\prec}(\frac{1}{N}),\end{aligned} \tag{4.13}$$

where we have used (4.5). Plugging the estimates in (4.13) into (4.12) and using the fact $g_{ii}, \widetilde{g}_{ii} \sim N_{\mathbb{C}}(0, N^{-1})$ again, we can get the bounds

$$|\delta_{1i}| \prec \frac{1}{\sqrt{N}}\,, \qquad |\delta_{2i}| \prec \frac{1}{\sqrt{N}}. \tag{4.14}$$

Denote

$$\Delta_i := \boldsymbol{w}_i\boldsymbol{w}_i^* - \boldsymbol{r}_i\boldsymbol{r}_i^*.$$

Fix now $z \in \mathbb{C}^+$. Dropping z from the notation, a first order Neumann expansion of the resolvent yields

$$G_{kj} = G^{(i)}_{kj} - \left(G(\Delta_i\widetilde{B}^{\langle i\rangle}W_i + W_i\widetilde{B}^{\langle i\rangle}\Delta_i + \Delta_i\widetilde{B}^{\langle i\rangle}\Delta_i)G^{(i)}\right)_{kj}. \tag{4.15}$$

Observe that the second term on the right side of (4.15) is a polynomial in the terms

$$\begin{array}{llll} G^{(i)}_{ij}\,, & \boldsymbol{g}_i^*G^{(i)}\boldsymbol{e}_j & \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_j\,, & \boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_j\,,\\ G_{ki}\,, & \boldsymbol{e}_k^*G\boldsymbol{g}_i\,, & \boldsymbol{e}_k^*G\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\,, & \boldsymbol{e}_k^*G\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i\,,\\ \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i\,, & \boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\,, & \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i, & \boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i\,, \end{array} \tag{4.16}$$

with coefficients of the form $\delta_{1i}^{k_1}\delta_{2i}^{k_2}$, for some nonnegative integers k_1, k_2 such that $k_1 + k_2 \geq 1$. By assumption (4.10), the fact $\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i = b_i\boldsymbol{e}_i$, and assumption (2.12), we further observe that the first four terms in (4.16) are stochastically dominated by one. The last four terms are also stochastically dominated by one as follows from the trivial fact $\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i = b_i$ and Lemma 3.1. The terms in the second line of (4.16) are stochastically dominated by

$$|\boldsymbol{e}_k^*GQ^{\langle i\rangle}\boldsymbol{x}_i| \prec \|Q^{\langle i\rangle}\|\|G^*\boldsymbol{e}_k\|_2 \lesssim \sqrt{(GG^*)_{kk}} = \sqrt{\frac{\operatorname{Im}G_{kk}}{\eta}} \prec \frac{1}{\sqrt{\eta}}\,, \tag{4.17}$$

with $Q^{\langle i\rangle} = I$ or $\widetilde{B}^{\langle i\rangle}$, and with $\boldsymbol{x}_i = \boldsymbol{e}_i$ or $\boldsymbol{g}_i$, where the last step follows from (4.10). Note that the terms in the second line of (4.16) appear only linearly in (4.15). Hence, (4.14), (4.17) and the order one bound for the first and last four terms in (4.16) lead to (4.11). □

5. Concentration with Respect to the Vector $\boldsymbol{g_i}$

In this section, we show that $G^{(i)}_{ii}$ concentrates around the partial expectation $\mathbb{E}_{\boldsymbol{g}_i}[G^{(i)}_{ii}]$, where $\mathbb{E}_{\boldsymbol{g}_i}[\,\cdot\,]$ is the expectation with respect to the collection $(\operatorname{Re} g_{ij}, \operatorname{Im} g_{ij})_{j=1}^{N}$. Besides the diagonal Green function entries $G^{(i)}_{ii} = \boldsymbol{e}_i^* G^{(i)} \boldsymbol{e}_i$ the following combinations are of importance

$$T_i(z) := \boldsymbol{g}_i^* G^{(i)}(z) \boldsymbol{e}_i\,, \qquad S_i(z) := \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_i\,, \qquad z \in \mathbb{C}^+. \tag{5.1}$$

The estimation of $\mathbb{E}_{\boldsymbol{g}_i}[G^{(i)}_{ii}]$, carried out in the Sects. 6 and 7, involves the quantities T_i and S_i. From a technical point of view, it is convenient to be able to go back and forth between T_i, S_i and their expectations $\mathbb{E}_{\boldsymbol{g}_i}[T_i]$, $\mathbb{E}_{\boldsymbol{g}_i}[S_i]$. Thus after establishing concentration estimates for $G^{(i)}_{ii}$ in Lemma 5.1 below, we establish in Corollary 5.2 concentration estimates for T_i and S_i where we also give a rough bounds on T_i, S_i and related quantities. We need some more notation: for a general random variable X we define

$$\mathbb{E}_{\boldsymbol{g}_i} X := X - \mathbb{E}_{\boldsymbol{g}_i} X. \tag{5.2}$$

The main task in this section is to prove the following lemma.

Lemma 5.1. *Suppose that the assumptions of Theorem 2.5 are satisfied and let $\gamma > 0$. Fix $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ and assume that*

$$\left|G_{ii}(z) - (a_i - \omega_B(z))^{-1}\right| \prec N^{-\frac{\gamma}{4}}\,, \qquad \left|G^{(i)}_{ii}(z) - (a_i - \omega_B(z))^{-1}\right| \prec N^{-\frac{\gamma}{4}}\,, \tag{5.3}$$

uniformly in $i \in [\![1, N]\!]$. Then

$$\max_{i\in[\![1,N]\!]} \left|\mathbb{E}_{\boldsymbol{g}_i}[G^{(i)}_{ii}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{5.4}$$

Proof of Lemma 5.1. In this proof we fix $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. Recall the definition of $G^{\langle i\rangle}(z)$ in (4.4) and note that $G^{\langle i\rangle}(z)$ is independent of $\boldsymbol{v}_i$ (or $\boldsymbol{g}_i$). It is therefore natural to expand $G^{(i)}(z)$ around $G^{\langle i\rangle}(z)$ and to use the independence between $G^{\langle i\rangle}(z)$ and $\boldsymbol{g}_i$ in order to verify the concentration estimates. However, by construction, we have

$$G^{\langle i\rangle}_{ii}(z) = \frac{1}{a_i + b_i - z}\,, \tag{5.5}$$

which may be as large as $1/\eta$, depending on a_i, b_i and z. To circumvent problems coming from instabilities in $G^{\langle i\rangle}_{ii}(z)$, we may use a "regularization" trick to enhance stability in the $\boldsymbol{e}_i$-direction: instead of considering the Green function of $H^{(i)} = A + \widetilde{B}^{(i)}$ directly, we first consider the (z-dependent) matrix

$$H^{\{i\}}(z) := A + \widetilde{B}^{(i)} - (b_i + \omega_B(z) - z)\boldsymbol{e}_i \boldsymbol{e}_i^*\,, \tag{5.6}$$

and define $G^{\{i\}}(z) := (H^{\{i\}}(z) - z)^{-1}$. Note that $H^{\{i\}}(z)$ is not symmetric, yet since $\operatorname{Im}\omega_B(z) \geq \operatorname{Im} z$ by Proposition 2.1, $G^{\{i\}}(z)$ is in fact well-defined on the whole upper-half plane. Fix any $j \in [\![1, N]\!]$. Using the rank-one perturbation formula (3.2), we get

$$G_{ij}^{(i)}(z) = G_{ij}^{\{i\}}(z) - \frac{(b_i + \omega_B(z) - z)G_{ii}^{\{i\}}(z)G_{ij}^{\{i\}}(z)}{1 + (b_i + \omega_B(z) - z)G_{ii}^{\{i\}}(z)} = \frac{G_{ij}^{\{i\}}(z)}{1 + (b_i + \omega_B(z) - z)G_{ii}^{\{i\}}(z)}. \tag{5.7}$$

Some algebra then reveals that

$$G_{ij}^{(i)}(z) = \frac{G_{ij}^{\{i\}}(z) - (b_i + \omega_B(z) - z)\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)]\, G_{ij}^{(i)}(z)}{1 + (b_i + \omega_B(z) - z)\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)]}. \tag{5.8}$$

By assumption (5.3) and identity (5.7), we have

$$\left|G_{ii}^{(i)}(z) - \frac{1}{a_i - \omega_B(z)}\right| \prec N^{-\frac{\gamma}{4}}, \qquad \left|G_{ii}^{\{i\}}(z) - \frac{1}{a_i - b_i - 2\omega_B(z) + z}\right| \prec N^{-\frac{\gamma}{4}}. \tag{5.9}$$

Note that $|\mathrm{Im}\,(a_i - b_i - 2\omega_B(z) + z)| \geq \mathrm{Im}\,\omega_B(z) > 0$, thus both denominators are well separated away from 0 by Lemma 3.4, in particular $G_{ii}^{(i)}(z)$ and $G_{ii}^{\{i\}}(z)$ are uniformly bounded. We will prove below that

$$\left|\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{5.10}$$

Setting $j = i$ in (5.8) and expressing the denominator on the right side by using (5.9)–(5.10), we get

$$\left|1 + (b_i + \omega_B(z) - z)\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)] - \frac{a_i - \omega_B(z)}{a_i - b_i - 2\omega_B(z) + z}\right| \prec N^{-\frac{\gamma}{4}}. \tag{5.11}$$

In particular, together with Lemma 3.4 and $\mathrm{Im}\,\omega_B(z) \geq \mathrm{Im}\,z$, this implies that the absolute value of the denominator on the right side of (5.8) is bounded from below by some strictly positive constant. Thus, applying $\mathbb{E}_{\boldsymbol{g}_i}$ on both sides of (5.10), we obtain the concentration estimate in (5.4).

In the rest of the proof, we verify (5.10). Consider next the matrix

$$H^{[i]}(z) := A + \widetilde{B}^{\langle i\rangle} - (b_i + \omega_B(z) - z)\boldsymbol{e}_i\boldsymbol{e}_i^*, \tag{5.12}$$

and let $G^{[i]}(z) := (H^{[i]}(z) - z)^{-1}$. Note that $H^{[i]}(z)$ depends on z and $\omega_B(z)$ and is thus not symmetric, yet since $\mathrm{Im}\,\omega_B(z) \geq \mathrm{Im}\,z$, $G^{[i]}(z)$ is, similar to $G^{\{i\}}$, well-defined on the whole upper-half plane. Note that

$$|G_{ii}^{[i]}(z)| = \left|\frac{1}{a_i - \omega_B(z)}\right| \lesssim 1 \tag{5.13}$$

since $\mathrm{Im}\,\omega_B$ is uniformly bounded from below on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ by Lemma 3.4.

We now expand $G^{\{i\}}(z)$ around $G^{[i]}(z)$ and use the independence among $G^{[i]}(z)$ and $\boldsymbol{g}_i$. For simplicity, we hereafter drop the z-dependence from the notation. We start with noticing that

$$\begin{aligned} H^{\{i\}} - H^{[i]} &= -\boldsymbol{w}_i\boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle} - \widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\boldsymbol{w}_i^* + \boldsymbol{w}_i\boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\boldsymbol{w}_i^* \\ &= -\boldsymbol{w}_i\boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle} - \left(\widetilde{B}^{\langle i\rangle} - \boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\ I\right)\boldsymbol{w}_i\boldsymbol{w}_i^* \\ &= \boldsymbol{w}_i\boldsymbol{s}_i^* + \boldsymbol{t}_i\boldsymbol{w}_i^*, \end{aligned} \tag{5.14}$$

where we introduced

$$\boldsymbol{s}_i := -\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\,, \qquad \boldsymbol{t}_i := -\big(\widetilde{B}^{\langle i\rangle} - \boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\ I\big)\boldsymbol{w}_i. \tag{5.15}$$

Iterating the rank-one perturbation formula (3.2) once, we obtain

$$G^{\{i\}} = X^{[i]} - \frac{X^{[i]}\boldsymbol{t}_i\boldsymbol{w}_i^*X^{[i]}}{1+\boldsymbol{w}_i^*X^{[i]}\boldsymbol{t}_i}\,, \qquad X^{[i]} := G^{[i]} - \frac{G^{[i]}\boldsymbol{w}_i\boldsymbol{s}_i^*G^{[i]}}{1+\boldsymbol{s}_i^*G^{[i]}\boldsymbol{w}_i}. \tag{5.16}$$

Substituting the second identity in (5.16) to the first one, we obtain

$$\begin{aligned} G^{\{i\}} = G^{[i]} &+ \frac{\boldsymbol{s}_i^*G^{[i]}\boldsymbol{t}_i}{1+\Xi_i}G^{[i]}\boldsymbol{w}_i\boldsymbol{w}_i^*G^{[i]} + \frac{\boldsymbol{w}_i^*G^{[i]}\boldsymbol{w}_i}{1+\Xi_i}G^{[i]}\boldsymbol{t}_i\boldsymbol{s}_i^*G^{[i]} \\ &- \frac{1+\boldsymbol{s}_i^*G^{[i]}\boldsymbol{w}_i}{1+\Xi_i}G^{[i]}\boldsymbol{t}_i\boldsymbol{w}_i^*G^{[i]} - \frac{1+\boldsymbol{w}_i^*G^{[i]}\boldsymbol{t}_i}{1+\Xi_i}G^{[i]}\boldsymbol{w}_i\boldsymbol{s}_i^*G^{[i]}, \end{aligned} \tag{5.17}$$

where

$$\Xi_i := (\boldsymbol{s}_i^*G^{[i]}\boldsymbol{w}_i) + (\boldsymbol{w}_i^*G^{[i]}\boldsymbol{t}_i) + (\boldsymbol{s}_i^*G^{[i]}\boldsymbol{w}_i)(\boldsymbol{w}_i^*G^{[i]}\boldsymbol{t}_i) - (\boldsymbol{w}_i^*G^{[i]}\boldsymbol{w}_i)(\boldsymbol{s}_i^*G^{[i]}\boldsymbol{t}_i). \tag{5.18}$$

Taking the $(i,\,j)$th matrix entry in (5.17), we get

$$G_{ij}^{\{i\}} = G_{ij}^{[i]} + \frac{\Psi_{i,j}}{1+\Xi_i}\,, \tag{5.19}$$

where

$$\begin{aligned} \Psi_{i,j} := &-(\boldsymbol{e}_i^*G^{[i]}\boldsymbol{t}_i)\big((\boldsymbol{w}_i^*G^{[i]}\boldsymbol{e}_j) + (\boldsymbol{s}_i^*G^{[i]}\boldsymbol{w}_i)(\boldsymbol{w}_i^*G^{[i]}\boldsymbol{e}_j) - (\boldsymbol{w}_i^*G^{[i]}\boldsymbol{w}_i)(\boldsymbol{s}_i^*G^{[i]}\boldsymbol{e}_j)\big) \\ &- (\boldsymbol{e}_i^*G^{[i]}\boldsymbol{w}_i)\big((\boldsymbol{s}_i^*G^{[i]}\boldsymbol{e}_j) + (\boldsymbol{w}_i^*G^{[i]}\boldsymbol{t}_i)(\boldsymbol{s}_i^*G^{[i]}\boldsymbol{e}_j) - (\boldsymbol{s}_i^*G^{[i]}\boldsymbol{t}_i)(\boldsymbol{w}_i^*G^{[i]}\boldsymbol{e}_j)\big). \end{aligned} \tag{5.20}$$

We now rewrite (5.19) as

$$G_{ij}^{\{i\}} = G_{ij}^{[i]} + \frac{\Psi_{i,j} - \mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]\,(G_{ii}^{\{i\}} - G_{ij}^{[i]})}{1+\mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]}. \tag{5.21}$$

Since $|G_{ii}^{\{i\}}| \prec 1$ (*c.f.* (5.9)) and $|G_{ii}^{[i]}| \prec 1$ (*c.f.* (5.13)), it suffices to verify the following statements to show (5.10):

$$\text{(i)}\quad \big|\mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]\big| \prec \frac{1}{\sqrt{N\eta}}\,, \quad \text{(ii)}\quad \frac{1}{1+\mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]} \prec 1\,, \quad \text{(iii)}\quad \big|\mathbb{E}_{\boldsymbol{g}_i}[\Psi_{i,j}]\big| \prec \frac{1}{\sqrt{N\eta}}. \tag{5.22}$$

We first show claim (i). Substituting the definitions in (5.15) into (5.18), we have

$$\begin{aligned} \Xi_i = &-\boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}G^{[i]}\boldsymbol{w}_i - \boldsymbol{w}_i^*G^{[i]}\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i + \boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}G^{[i]}\boldsymbol{w}_i\ \boldsymbol{w}_i^*G^{[i]}\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i \\ &- \boldsymbol{w}_i^*G^{[i]}\boldsymbol{w}_i\ \boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}G^{[i]}\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i + \boldsymbol{w}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{w}_i\ \boldsymbol{w}_i^*G^{[i]}\boldsymbol{w}_i. \end{aligned} \tag{5.23}$$

Let $Q_1^{\langle i\rangle}$ and $Q_2^{\langle i\rangle}$ each stand for either I or $\widetilde{B}^{\langle i\rangle}$. Recalling that $\boldsymbol{w}_i = \boldsymbol{e}_i + \boldsymbol{g}_i$ and that $\boldsymbol{g}_i \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1}I)$ is a complex Gaussian vector, we compute

$$\mathbb{E}_{\boldsymbol{g}_i} \boldsymbol{w}_i^* Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle} \boldsymbol{w}_i = (Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle})_{ii} + \operatorname{tr} Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle}. \tag{5.24}$$

To bound the right side of (5.24) we observe that $|(Q_1 G^{[i]} Q_2)_{ii}| \prec |G_{ii}^{[i]}| \prec 1$, where we used that $\boldsymbol{e}_i$ is an eigenvector of $\widetilde{B}^{\langle i\rangle}$ and (5.13). (Notice that, for simplicity, here and at several other places we consistently use the notation $\prec$ even when the stronger $\leq$ or $\lesssim$ relations would also hold, *i.e.* we use the concept stochastic domination even for estimating almost surely bounded or deterministic quantities.)

To control the second term on the right side of (5.24), we note that a first order Neumann expansion of the resolvents yields

$$\begin{aligned}
|\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{[i]} - \operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}| &= |\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{[i]} (b_i + \omega_B(z) - z) \boldsymbol{e}_i \boldsymbol{e}_i^* G^{\langle i\rangle}| \\
&\prec \frac{1}{N} \|Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{[i]} \boldsymbol{e}_i \boldsymbol{e}_i^*\|_2 \|\boldsymbol{e}_i \boldsymbol{e}_i^* G^{\langle i\rangle}\|_2 \\
&\prec \frac{1}{N} \|G^{[i]} \boldsymbol{e}_i \boldsymbol{e}_i^*\|_2 \|\boldsymbol{e}_i \boldsymbol{e}_i^* G^{\langle i\rangle}\|_2 \\
&\prec \frac{1}{N} |(\boldsymbol{e}_i^* |G^{[i]}|^2 \boldsymbol{e}_i)|^{1/2} \, |(\boldsymbol{e}_i^* |G^{\langle i\rangle}|^2 \boldsymbol{e}_i)|^{1/2},
\end{aligned} \tag{5.25}$$

where we used the boundedness of b_i, $\omega_B(z)$, $\|Q_1^{\langle i\rangle}\|$ and $\|Q_2^{\langle i\rangle}\|$. Notice next the identities

$$(|G^{[i]}(z)|^2)_{jj} = \frac{\operatorname{Im} G_{jj}^{[i]}(z)}{(1-\delta_{ij})\eta + \delta_{ij} \operatorname{Im} \omega_B(z)}, \qquad (|G^{\langle i\rangle}(z)|^2)_{jj} = \frac{\operatorname{Im} G_{jj}^{\langle i\rangle}(z)}{\eta}, \tag{5.26}$$

for $j \in [\![1, N]\!]$, with $z = E + \mathrm{i}\eta$ and $|G|^2 = G^* G$. The second identity in (5.26) is the Ward identity that is valid for the Green function of any self-adjoint operator and it can be checked by spectral calculus. For the first identity in (5.26), recalling the definition in (5.12) and that $\boldsymbol{e}_j^*(A + \widetilde{B}^{\langle i\rangle})\boldsymbol{e}_i = (a_i + b_i)\delta_{ij}$, one sees that for any fixed i,

$$G_{ii}^{[i]}(z) = \frac{1}{a_i - \omega_B(z)}, \qquad G_{ij}^{[i]}(z) = 0 \qquad \forall j \neq i.$$

This implies $|G^{[i]}|_{ii}^2 = \sum_j G_{ij}^{[i]} (G^{[i]})_{ji}^* = |G_{ii}^{[i]}|^2$ thus the first identity in (5.26) with $j = i$ follows. For $j \neq i$, one can see the first identity of (5.26) by applying the Ward identity to the minor of $G^{[i]}$, with ith row and ith column removed. Since $|G_{ii}^{\langle i\rangle}| \prec \frac{1}{\eta}$ (*c.f.* (5.5)), we obtain combining (5.25) and (5.26) with (5.3) that

$$|\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{[i]} - \operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}| \prec \frac{1}{N\eta}. \tag{5.27}$$

Since $H^{\langle i\rangle}$ is a Hermitian finite-rank perturbation of H, we can apply (3.3) to conclude that

$$|\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G - \operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}| \prec \frac{1}{N\eta}. \tag{5.28}$$

We will now show that $\operatorname{tr} Q_1^{\langle i\rangle} G^{\langle i\rangle} Q_2^{\langle i\rangle}$ is bounded. Using the resolvent identities and $\operatorname{tr} B^{\langle i\rangle} = \operatorname{tr} B = 0$, we get

$$\operatorname{tr} B^{\langle i\rangle} G^{\langle i\rangle} = 1 - \operatorname{tr}(A-z)G^{\langle i\rangle}, \qquad \operatorname{tr} B^{\langle i\rangle} G^{\langle i\rangle} B^{\langle i\rangle} = z + \operatorname{tr}(A-z)G^{\langle i\rangle}(A-z),$$

thus to control $\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}$ we need to bound $\operatorname{tr}(A-z)^k G^{\langle i\rangle}$ for $k = 0, 1, 2$. Since $H^{\langle i\rangle}$ is a Hermitian finite-rank perturbation of H, we can apply (3.3) to conclude that

$$|\operatorname{tr}(A-z)^k G^{\langle i\rangle} - \operatorname{tr}(A-z)^k G| \prec \frac{1}{N\eta}, \qquad k = 0, 1, 2. \tag{5.29}$$

Since A is diagonal with bounded matrix elements, we have $\operatorname{tr}(A-z)^k G \lesssim \max_j |G_{jj}| \prec 1$, where the last bound comes from (5.3). This directly controls $\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}$ and then, using (5.27) and (5.28), we have

$$|\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{\langle i\rangle}| + |\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G| + |\operatorname{tr} Q_2^{\langle i\rangle} Q_1^{\langle i\rangle} G^{[i]}| \prec 1. \tag{5.30}$$

Thus, returning to (5.24), we showed

$$\mathbb{E}_{\boldsymbol{g}_i} \boldsymbol{w}_i^* Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle} \boldsymbol{w}_i \prec 1. \tag{5.31}$$

Using the Gaussian concentration estimates in (3.1) and $\boldsymbol{w}_i = \boldsymbol{e}_i + \boldsymbol{g}_i$, we obtain

$$\begin{aligned} |\mathbb{E}_{\boldsymbol{g}_i} \boldsymbol{w}_i^* Q_1^{\langle i\rangle} G^{[i]}(z) Q_2^{\langle i\rangle} \boldsymbol{w}_i| &\prec \left(\frac{(|Q_1^{\langle i\rangle} G^{[i]}(z) Q_2^{\langle i\rangle}|^2)_{ii}}{N}\right)^{\frac{1}{2}} + \left(\frac{\|Q_1^{\langle i\rangle} G^{[i]}(z) Q_2^{\langle i\rangle}\|_2^2}{N^2}\right)^{\frac{1}{2}} \\ &\prec \left(\frac{\operatorname{Im} G_{ii}^{[i]}(z)}{N \operatorname{Im} \omega_B(z)}\right)^{\frac{1}{2}} + \left(\frac{\operatorname{Im}\operatorname{tr} G^{[i]}(z)}{N\eta}\right)^{\frac{1}{2}} \prec \frac{1}{\sqrt{N\eta}}, \end{aligned} \tag{5.32}$$

where we also used that $\boldsymbol{e}_i$ is an eigenvector of $\widetilde{B}^{\langle i\rangle}$, that $\widetilde{B}^{\langle i\rangle}$ is bounded and (5.26). In the last step (5.13) and (5.30) were used. Combined with (5.31) we thus proved

$$\boldsymbol{w}_i^* Q_1^{\langle i\rangle} G^{[i]}(z) Q_2^{\langle i\rangle} \boldsymbol{w}_i \prec 1. \tag{5.33}$$

For a later use we remark that, combining (5.28) and (5.32), we also proved

$$\boldsymbol{w}_i^* Q_1^{\langle i\rangle} G^{[i]}(z) Q_2^{\langle i\rangle} \boldsymbol{w}_i = (Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle})_{ii} + \operatorname{tr} Q_1^{\langle i\rangle} G Q_2^{\langle i\rangle} + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N\eta}}\Big). \tag{5.34}$$

In a very similar way we get, recalling that $\operatorname{tr} B = 0$ and $\|B\| \prec 1$, that

$$\boldsymbol{w}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{w}_i = b_i + \mathbb{E}_{\boldsymbol{g}_i} \boldsymbol{w}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{w}_i = b_i + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N}}\Big). \tag{5.35}$$

To deal with terms containing four or six factors of $\boldsymbol{w}_i$ in $\mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]$ (see (5.23)), we use the following rough bound. For general random variables X and Y satisfying $|X|, |Y| \prec 1$, we have

$$\mathbb{E}_{\boldsymbol{g}_i}[XY] = \mathbb{E}_{\boldsymbol{g}_i}[\mathbb{E}_{\boldsymbol{g}_i}[X]\mathbb{E}_{\boldsymbol{g}_i}[Y]] + \mathbb{E}_{\boldsymbol{g}_i}[\mathbb{E}_{\boldsymbol{g}_i}[X]\mathbb{E}_{\boldsymbol{g}_i}[Y]] + \mathbb{E}_{\boldsymbol{g}_i}[\mathbb{E}_{\boldsymbol{g}_i}[X]\mathbb{E}_{\boldsymbol{g}_i}[Y]]. \tag{5.36}$$

In particular we have $|\mathbb{E}_{\boldsymbol{g}_i}[XY]| \prec |\mathbb{E}_{\boldsymbol{g}_i} X| + |\mathbb{E}_{\boldsymbol{g}_i} Y|$, where we used basic properties of stochastic domination outlined in Subsect. 3.1.

Then, recalling the explicit form of Ξ_i in (5.23) and using (5.32), (5.33), (5.35) and (5.36), a straightforward estimate shows that $|\mathbb{E}_{\boldsymbol{g}_i} \Xi_i| \prec \frac{1}{\sqrt{N\eta}}$, and claim (i) in (5.22) is thus proved.

We next show statement (ii) of (5.22). To compute the expectation $\mathbb{E}_{\boldsymbol{g}_i}[\Xi_i]$, we are going to use the identities

$$\operatorname{tr} \widetilde{B} G = 1 - \operatorname{tr}(A - z)G\,, \qquad \operatorname{tr} \widetilde{B} G \widetilde{B} = z + \operatorname{tr}(A - z)G(A - z)\,, \tag{5.37}$$

that follow from $(H - z)G(z) = 1$ and $\operatorname{tr} A = \operatorname{tr} B = 0$. Invoking assumption (5.3) we have

$$\operatorname{tr}\big((A - z)^k G\big) = \frac{1}{N} \sum_{i=1}^{N} \frac{(a_i - z)^k}{a_i - \omega_B} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})\,,$$

with $k \in \mathbb{N}$. Recalling further the shorthand notation $m_{\boxplus} \equiv m_{\mu_A \boxplus \mu_B}$ and from (2.13) that

$$m_{\boxplus} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{a_i - \omega_B}\,, \tag{5.38}$$

we get from the above that

$$\begin{aligned} \operatorname{tr} G &= m_{\boxplus} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})\,, \\ \operatorname{tr}\big((A - z)G\big) &= 1 + (\omega_B - z)m_{\boxplus} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})\,, \\ \operatorname{tr}\big((A - z)^2 G\big) &= \omega_B - 2z + (\omega_B - z)^2 m_{\boxplus} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}}). \end{aligned} \tag{5.39}$$

Thus from (5.34) we obtain

$$\begin{aligned} \boldsymbol{w}_i^* G^{[i]} \boldsymbol{w}_i &= m_{\boxplus} + \frac{1}{a_i - \omega_B} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})\,, \\ \boldsymbol{w}_i^* \widetilde{B}^{\langle i \rangle} G^{[i]} \boldsymbol{w}_i &= -(\omega_B - z)m_{\boxplus} + \frac{b_i}{a_i - \omega_B} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})\,, \\ \boldsymbol{w}_i^* G^{[i]} \widetilde{B}^{\langle i \rangle} \boldsymbol{w}_i &= -(\omega_B - z)m_{\boxplus} + \frac{b_i}{a_i - \omega_B} + (N^{-\frac{\gamma}{4}})\,, \\ \boldsymbol{w}_i^* \widetilde{B}^{\langle i \rangle} G^{[i]} \widetilde{B}^{\langle i \rangle} \boldsymbol{w}_i &= (\omega_B - z) + (\omega_B - z)^2 m_{\boxplus} + \frac{b_i^2}{a_i - \omega_B} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}}). \end{aligned} \tag{5.40}$$

Plugging (5.40) into (5.23), using the identity $\omega_A + \omega_B = z - 1/m_{\boxplus}$ and taking the expectation, a straightforward computation shows that

$$1 + \mathbb{E}_{\boldsymbol{g}_i}[\Xi_i] = \frac{(\omega_A - b_i)(2\omega_B - a_i + b_i - z)m_{\boxplus}}{a_i - \omega_B} + \mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}}). \tag{5.41}$$

Then from Lemma 3.4 one observes that statement (ii) of (5.22) holds. In fact, the first term on the right side of (5.41) is bounded away from zero uniformly on $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$.

We move on to statement (iii) of (5.22). Let $Q_1^{\langle i \rangle}$ and $Q_2^{\langle i \rangle}$ each stand again for either I or $\widetilde{B}^{\langle i \rangle}$. Then we note that

$$\boldsymbol{e}_i^* Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle} \boldsymbol{w}_i = (Q_1^{\langle i\rangle} G^{[i]} Q_2^{\langle i\rangle})_{ii} + \mathrm{O}_\prec\Big(\frac{1}{\sqrt{N}}\Big), \tag{5.42}$$

as follows from the Gaussian large deviation estimates in (3.1), assumption (5.3) and the fact that $\boldsymbol{e}_i$ is an eigenvector of $Q_1^{\langle i\rangle}$, $Q_2^{\langle i\rangle}$ and $G^{[i]}$. Having established (5.42), it suffices to recall (5.34) and (5.35) to conclude that $|\mathbb{E}_{\boldsymbol{g}_i}[\Psi_{i,i}]| \prec \frac{1}{\sqrt{N\eta}}$. This proves claim (iii) in (5.22) and thus completes the proof of Lemma 5.1. □

Corollary 5.2. *Suppose that the assumptions of Theorem 2.5 are satisfied and let $\gamma > 0$. Fix $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ and assume that*

$$\big|G_{ii}^{(i)}(z) - (a_i - \omega_B(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}, \quad \big|G_{ii}(z) - (a_i - \omega_B(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}, \tag{5.43}$$

hold for all $i \in [\![1, N]\!]$. Letting $Q_1^{\langle i\rangle}$, $Q_2^{\langle i\rangle}$ stand for I or $\widetilde{B}^{\langle i\rangle}$, and letting $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ stand for $\boldsymbol{g}_i$ or $\boldsymbol{e}_i$, we have the bound

$$\max_{i\in[\![1,N]\!]} \big|\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{(i)}(z) Q_2^{\langle i\rangle} \boldsymbol{y}_i\big| \prec 1. \tag{5.44}$$

In particular, $|S_i(z)|, |T_i(z)| \prec 1$, for all $i \in [\![1, N]\!]$. Moreover, we have

$$\max_{i\in[\![1,N]\!]} \big|\mathbb{E}_{\boldsymbol{g}_i}[T_i(z)]\big| \prec \frac{1}{\sqrt{N\eta}}, \qquad \max_{i\in[\![1,N]\!]} \big|\mathbb{E}_{\boldsymbol{g}_i}[S_i(z)]\big| \prec \frac{1}{\sqrt{N\eta}}. \tag{5.45}$$

Proof. Using once more (3.2), we can write

$$\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{(i)} Q_2^{\langle i\rangle} \boldsymbol{y}_i = \boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} Q_2^{\langle i\rangle} \boldsymbol{y}_i - \frac{(b_i + \omega_B - z)\, \boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} \boldsymbol{e}_i\, \boldsymbol{e}_i^* G^{\{i\}} Q_2^{\langle i\rangle} \boldsymbol{y}_i}{1 + (b_i + \omega_B - z) G_{ii}^{\{i\}}}.$$

Hence to prove the bound in (5.44) it suffices to bound $\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} Q_2^{\langle i\rangle} \boldsymbol{y}_i$ and $\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} \boldsymbol{e}_i$ with the choices $Q_1^{\langle i\rangle}$, $Q_2^{\langle i\rangle} = I$ or $\widetilde{B}^{\langle i\rangle}$ and $\boldsymbol{x}_i$, $\boldsymbol{y}_i = \boldsymbol{g}_i$ or $\boldsymbol{e}_i$. To do so, we expand $G^{\{i\}}$ around $G^{[i]}$. It turns out that $\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} Q_2^{\langle i\rangle} \boldsymbol{y}_i$ and $\boldsymbol{x}_i^* Q_1^{\langle i\rangle} G^{\{i\}} \boldsymbol{e}_i$ both are of the form $\widetilde{\Psi}_i/(1+\Xi_i)$, where Ξ_i is given in (5.18) and $\widetilde{\Psi}_i$ is a polynomial of the quantities appearing in (5.34), (5.35) and (5.42). Then (i) and (ii) of (5.22) imply that $(1 + \Xi_i)^{-1} \prec 1$, which together with the bounds in (5.34) and (5.42) leads to the conclusion (5.44).

To prove (5.45), we follow, mutatis mutandis, the proof of (5.4) by replacing $G_{ii}^{(i)}$ by $T_i = \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i$ or $S_i = \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_i$. For instance, for T_i the counterpart of (5.8) is

$$\boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i = \frac{\boldsymbol{g}_i^* G^{\{i\}} \boldsymbol{e}_i - (b_i + \omega_B - z)\mathbb{E}_{\boldsymbol{g}_i}\big[G_{ii}^{\{i\}}\big]\, \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i}{1 + (b_i + \omega_B - z)\mathbb{E}_{\boldsymbol{g}_i}\big[G_{ii}^{\{i\}}\big]}.$$

Now, according to (5.11), (5.10) and the bound $|T_i| \prec 1$ (*c.f.* (5.44)), it suffices to show

$$\big|\mathbb{E}_{\boldsymbol{g}_i}\big[\boldsymbol{g}_i^* G^{\{i\}} \boldsymbol{e}_i\big]\big| \prec \frac{1}{\sqrt{N\eta}}. \tag{5.46}$$

The proof of (5.46) is nearly the same as the one of (5.10). One can also use a similar argument for S_i by using the bound $|S_i| \prec 1$ from (5.44). We omit the details. □

6. Identification of the Partial Expectation $\mathbb{E}_{\boldsymbol{g}_i}\big[G^{(i)}_{ii}\big]$

In this section, we estimate the partial expectation $\mathbb{E}_{\boldsymbol{g}_i}\big[G^{(i)}_{ii}\big]$, which together with the concentration inequalities in Lemma 5.1 lead to the following lemma. Recall the definition of S_i and T_i in (5.1).

Proposition 6.1. *Suppose that the assumptions of Theorem 2.5 are satisfied and let $\gamma > 0$. Fix $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. Assume that*

$$\Big|G^{(i)}_{ii}(z) - \big(a_i - \omega_B(z)\big)^{-1}\Big| \prec N^{-\frac{\gamma}{4}}\,, \quad \Big|G_{ii}(z) - \big(a_i - \omega_B(z)\big)^{-1}\Big| \prec N^{-\frac{\gamma}{4}}\,, \tag{6.1}$$

hold uniformly in $i \in [\![1, N]\!]$. Then,

$$\max_{i\in[\![1,N]\!]} \Big|G^{(i)}_{ii}(z) - \big(a_i - \omega^c_B(z)\big)^{-1}\Big| \prec \frac{1}{\sqrt{N\eta}}\,, \tag{6.2}$$

and

$$\max_{i\in[\![1,N]\!]} \Big|S_i(z) + \frac{z - \omega^c_B(z)}{a_i - \omega^c_B(z)}\Big| \prec \frac{1}{\sqrt{N\eta}}\,, \qquad \max_{i\in[\![1,N]\!]} \big|T_i(z)\big| \prec \frac{1}{\sqrt{N\eta}}. \tag{6.3}$$

In the proof of Proposition 6.1 we will need the following auxiliary lemma whose proof is postponed to the very end of this section.

Lemma 6.2. *Under the assumption of Proposition 6.1, the estimates*

$$\big|\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z) - \widetilde{B}G(z)\big)\big| \le \frac{C}{N\eta}\,, \qquad \big|\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z)\widetilde{B}^{\langle i\rangle} - \widetilde{B}G(z)\widetilde{B}\big)\big| \le \frac{C}{N\eta}\,, \tag{6.4}$$

and the bounds

$$\big|\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z)\big)\big| \prec 1\,, \qquad \big|\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z)\widetilde{B}^{\langle i\rangle}\big)\big| \prec 1\,, \tag{6.5}$$

hold uniformly in $i \in [\![1, N]\!]$. Furthermore the estimates

$$\big|\mathbb{E}_{\boldsymbol{g}_i}\big[\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z)\big)\big]\big| \le \frac{C}{N\eta}\,, \qquad \big|\mathbb{E}_{\boldsymbol{g}_i}\big[\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}(z)\widetilde{B}^{\langle i\rangle}\big)\big]\big| \le \frac{C}{N\eta}\,, \tag{6.6}$$

hold uniformly in $i \in [\![1, N]\!]$.

Proof of Proposition 6.1. Fix $i \in [\![1, N]\!]$. By the concentration results of Lemma 5.1 and Corollary 5.2, it suffices to estimate $\mathbb{E}_{\boldsymbol{g}_i}[G^{(i)}_{ii}(z)]$, $\mathbb{E}_{\boldsymbol{g}_i}\big[S_i(z)\big]$ and $\mathbb{E}_{\boldsymbol{g}_i}[T_i(z)]$ to establish (6.2) and (6.3). Recall the definition of $H^{(i)}$ and $G^{(i)}$ from (4.9). We start with the identity

$$(A - z)G^{(i)}(z) = -\widetilde{B}^{(i)}G^{(i)}(z) + I\,, \quad z \in \mathbb{C}^+. \tag{6.7}$$

Since A is diagonal, we have

$$(a_i - z)G^{(i)}_{ii}(z) = -\big(\widetilde{B}^{(i)}G^{(i)}(z)\big)_{ii} + 1\,, \quad z \in \mathbb{C}^+. \tag{6.8}$$

Therefore, to estimate $\mathbb{E}_{\boldsymbol{g}_i}[G^{(i)}_{ii}(z)]$, it suffices to estimate $\mathbb{E}_{\boldsymbol{g}_i}[(\widetilde{B}^{(i)}G^{(i)}(z))_{ii}]$ instead. Recalling the definitions in (4.7) and (4.8), we have

$$\begin{aligned}\big(\widetilde{B}^{(i)}G^{(i)}\big)_{ii} &= \boldsymbol{e}_i^*\big(I-\boldsymbol{e}_i\boldsymbol{e}_i^*-\boldsymbol{e}_i\boldsymbol{g}_i^*-\boldsymbol{g}_i\boldsymbol{e}_i^*-\boldsymbol{g}_i\boldsymbol{g}_i^*\big)\widetilde{B}^{\langle i\rangle}\\ &\quad\times\big(I-\boldsymbol{e}_i\boldsymbol{e}_i^*-\boldsymbol{e}_i\boldsymbol{g}_i^*-\boldsymbol{g}_i\boldsymbol{e}_i^*-\boldsymbol{g}_i\boldsymbol{g}_i^*\big)G^{(i)}\boldsymbol{e}_i\\ &= -\boldsymbol{e}_i^*\big(\boldsymbol{e}_i\boldsymbol{g}_i^*+\boldsymbol{g}_i\boldsymbol{e}_i^*+\boldsymbol{g}_i\boldsymbol{g}_i^*\big)\widetilde{B}^{\langle i\rangle}\big(I-\boldsymbol{e}_i\boldsymbol{e}_i^*-\boldsymbol{e}_i\boldsymbol{g}_i^*-\boldsymbol{g}_i\boldsymbol{e}_i^*-\boldsymbol{g}_i\boldsymbol{g}_i^*\big)G^{(i)}\boldsymbol{e}_i.\end{aligned} \tag{6.9}$$

Since $\boldsymbol{e}_i$ is an eigenvector of $\widetilde{B}^{\langle i\rangle}$ (*c.f.* (4.3)), we have $(\widetilde{B}^{\langle i\rangle}G^{(i)})_{ii} = b_i G^{(i)}_{ii}$. Since moreover B is traceless by assumption (2.27), we have $\operatorname{tr}\widetilde{B}^{\langle i\rangle} = \operatorname{tr} B = 0$. Thus the a priori estimates in (6.1), the bound in (5.44), and the following concentration estimates (*c.f.* Lemma 3.1)

$$|\boldsymbol{e}_j^*\boldsymbol{g}_i| \prec \frac{1}{\sqrt{N}}, \qquad |\boldsymbol{e}_j^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i| \prec \frac{1}{\sqrt{N}}, \qquad \big|\boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\big| \prec \frac{1}{\sqrt{N}}, \tag{6.10}$$

for all $j \in [\![1, N]\!]$, imply that $\boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_i$ is the only relevant term in (6.9). Thus recalling from definition (5.1) that $S_i = \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_i$ we arrive at

$$\big|(\widetilde{B}^{(i)}G^{(i)})_{ii} + S_i\big| \prec \frac{1}{\sqrt{N}}. \tag{6.11}$$

Using integration by parts for complex Gaussian random variables, we compute $\mathbb{E}_{\boldsymbol{g}_i}[S_i]$ next. Regarding g and $\overline{g}$ as independent variables for computing $\partial_g f(g,\overline{g})$, we have

$$\int_{\mathbb{C}} \overline{g}\, f(g,\overline{g})\, \mathrm{e}^{-\frac{|g|^2}{\sigma^2}}\,\mathrm{d}g\wedge\mathrm{d}\overline{g} = \sigma^2\int_{\mathbb{C}} \partial_g f(g,\overline{g})\,\mathrm{e}^{-\frac{|g|^2}{\sigma^2}}\,\mathrm{d}g\wedge\mathrm{d}\overline{g}, \tag{6.12}$$

for differentiable functions $f : \mathbb{C}^2 \to \mathbb{C}$. Using (6.12) with $\sigma^2 = 1/N$ for each component of $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{iN})$, we have

$$\mathbb{E}_{\boldsymbol{g}_i}[S_i] = \sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{g}_i}\big[\overline{g}_{ik}(\widetilde{B}^{\langle i\rangle}G^{(i)})_{ki}\big] = \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\boldsymbol{g}_i}\bigg[\frac{\partial(\widetilde{B}^{\langle i\rangle}G^{(i)})_{ki}}{\partial g_{ik}}\bigg]. \tag{6.13}$$

Using the definitions in (4.7), (4.8) and regarding $g_{ik}, \overline{g}_{ik}$ as independent variables, we have

$$\frac{\partial W_i}{\partial g_{ik}} = -\boldsymbol{e}_k\boldsymbol{e}_i^* - \boldsymbol{e}_k\boldsymbol{g}_i^*, \tag{6.14}$$

so that

$$\begin{aligned}\frac{\partial\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\big)_{ki}}{\partial g_{ik}} &= \boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G^{(i)}\big(\boldsymbol{e}_k\boldsymbol{e}_i^*+\boldsymbol{e}_k\boldsymbol{g}_i^*\big)\widetilde{B}^{\langle i\rangle}\big(I-\boldsymbol{e}_i\boldsymbol{e}_i^*-\boldsymbol{e}_i\boldsymbol{g}_i^*-\boldsymbol{g}_i\boldsymbol{e}_i^*-\boldsymbol{g}_i\boldsymbol{g}_i^*\big)G^{(i)}\boldsymbol{e}_i\\ &\quad+\boldsymbol{e}_k^*\widetilde{B}^{\langle i\rangle}G^{(i)}\big(I-\boldsymbol{e}_i\boldsymbol{e}_i^*-\boldsymbol{e}_i\boldsymbol{g}_i^*-\boldsymbol{g}_i\boldsymbol{e}_i^*-\boldsymbol{g}_i\boldsymbol{g}_i^*\big)\widetilde{B}^{\langle i\rangle}\big(\boldsymbol{e}_k\boldsymbol{e}_i^*+\boldsymbol{e}_k\boldsymbol{g}_i^*\big)G^{(i)}\boldsymbol{e}_i.\end{aligned} \tag{6.15}$$

Since $\boldsymbol{e}_i$ is an eigenvector of $\widetilde{B}^{\langle i\rangle}$ with eigenvalue b_i, we further get from (6.15) that

$$\begin{aligned}\frac{\partial\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\big)_{ki}}{\partial g_{ik}} &= \big(\widetilde{B}^{\langle i\rangle}G^{(i)}\big)_{kk}\big(\boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_i - b_i\,\boldsymbol{g}_i^*G^{(i)}\boldsymbol{e}_i\big)\\ &\quad+\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\widetilde{B}^{\langle i\rangle}\big)_{kk}\big(G^{(i)}_{ii}+\boldsymbol{g}_i^*G^{(i)}\boldsymbol{e}_i\big)\end{aligned}$$

$$
\begin{aligned}
&- \big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)_{kk} \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big) \big(\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i + \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_i + \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i\big) \\
&- \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\big(\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k\big)\big(\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_i\big) \\
&- \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\big(\boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k\big)\big(\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_i\big) \\
&- \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\big(\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k\big)\big(\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{g}_i\big) \\
&- \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\big(\boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_k\big)\big(\boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{g}_i\big). \qquad (6.16)
\end{aligned}
$$

Plugging (6.16) into (6.13) and rearranging, we get

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}[S_i] &= \mathbb{E}_{\boldsymbol{g}_i}\Big[\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)\big(\boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_i - b_i\ \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\Big] \\
&\quad + \mathbb{E}_{\boldsymbol{g}_i}\Big[\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)} \widetilde{B}^{\langle i\rangle}\big)\big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\Big] \\
&\quad - \mathbb{E}_{\boldsymbol{g}_i}\Big[\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)\ \big(\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i + \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{e}_i + \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i\big)\big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\Big] \\
&\quad - \frac{1}{N}\mathbb{E}_{\boldsymbol{g}_i}\Big[\Big(b_i^2 G^{(i)}_{ii} + \boldsymbol{g}_i^*\big(\widetilde{B}^{\langle i\rangle}\big)^2 G^{(i)} \boldsymbol{e}_i + \boldsymbol{e}_i^*(\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{g}_i + \boldsymbol{g}_i^*\big(\widetilde{B}^{\langle i\rangle}\big)^2 G^{(i)} \boldsymbol{g}_i\Big) \\
&\qquad\qquad \times \big(G^{(i)}_{ii} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i\big)\Big]. \qquad (6.17)
\end{aligned}
$$

We next claim that the last two terms on the right of (6.17) are small. Using the boundedness of $G^{(i)}_{ii}$ (following from the a priori estimate (6.1)), the bound (5.44), the concentration estimates in (6.10), and estimate (6.5) of the auxiliary Lemma 6.2, and the trivial bounds

$$
\big|\boldsymbol{x}_i^*(\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{y}_i\big| \prec \frac{1}{\eta}, \qquad \boldsymbol{x}_i,\ \boldsymbol{y}_i = \boldsymbol{e}_i \ \text{ or } \ \boldsymbol{g}_i, \qquad (6.18)
$$

we see that the last two terms on the right side of (6.17) are indeed negligible, *i.e.*

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}[S_i] &= \mathbb{E}_{\boldsymbol{g}_i}\Big[\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)\big(S_i - b_i T_i\big)\Big] \\
&\quad + \mathbb{E}_{\boldsymbol{g}_i}\Big[\operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)} \widetilde{B}^{\langle i\rangle}\big)\big(G^{(i)}_{ii} + T_i\big)\Big] + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N}}\Big) + \mathrm{O}_{\prec}\Big(\frac{1}{N\eta}\Big), \qquad (6.19)
\end{aligned}
$$

where we also used the definitions of T_i and S_i in (5.1). From assumption (6.1) and Corollary 5.2, we have the bounds

$$
\max_{j\in[\![1,N]\!]} |G^{(j)}_{jj}| \prec 1, \qquad \max_{j\in[\![1,N]\!]} |T_j| \prec 1, \qquad \max_{j\in[\![1,N]\!]} |S_j| \prec 1. \qquad (6.20)
$$

We hence obtain from (6.19), (6.5), and the concentration estimates in (6.6), (5.4) that

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}\big[S_i\big] &= \operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)\ \big(\mathbb{E}_{\boldsymbol{g}_i}\big[S_i\big] - b_i\, \mathbb{E}_{\boldsymbol{g}_i}\big[T_i\big]\big) \\
&\quad + \operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)} \widetilde{B}^{\langle i\rangle}\big)\ \big(G^{(i)}_{ii} + \mathbb{E}_{\boldsymbol{g}_i}\big[T_i\big]\big) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big). \qquad (6.21)
\end{aligned}
$$

Repeating the above computations for $\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_i] = \mathbb{E}_{\boldsymbol{g}_i}[T_i]$, we similarly obtain

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}\big[T_i\big] &= \operatorname{tr} G^{(i)}\ \big(\mathbb{E}_{\boldsymbol{g}_i}\big[S_i\big] - b_i\, \mathbb{E}_{\boldsymbol{g}_i}\big[T_i\big]\big) \\
&\quad + \operatorname{tr}\big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)\ \big(G^{(i)}_{ii} + \mathbb{E}_{\boldsymbol{g}_i}\big[T_i\big]\big) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big). \qquad (6.22)
\end{aligned}
$$

Now, using the bounds in (6.20), the estimates (6.4) and $|\operatorname{tr} G^{(i)} - \operatorname{tr} G| \prec \frac{1}{N\eta}$ (following from (3.3)), we obtain from (6.21) and (6.22) the equations

$$\mathbb{E}_{\boldsymbol{g}_i}[S_i] - \operatorname{tr}\left(\widetilde{B}G\widetilde{B}\right)\left(G_{ii}^{(i)} + \mathbb{E}_{\boldsymbol{g}_i}[T_i]\right) = \operatorname{tr}\left(\widetilde{B}G\right)\left(\mathbb{E}_{\boldsymbol{g}_i}[S_i] - b_i\mathbb{E}_{\boldsymbol{g}_i}[T_i]\right) + \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right), \tag{6.23}$$

and

$$\mathbb{E}_{\boldsymbol{g}_i}[T_i] - \operatorname{tr}\left(\widetilde{B}G\right)\left(G_{ii}^{(i)} + \mathbb{E}_{\boldsymbol{g}_i}[T_i]\right) = \operatorname{tr}\left(G\right)\left(\mathbb{E}_{\boldsymbol{g}_i}[S_i] - b_i\mathbb{E}_{\boldsymbol{g}_i}[T_i]\right) + \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right). \tag{6.24}$$

We first approximately solve (6.24) for $\mathbb{E}_{\boldsymbol{g}_i}[T_i]$ to show, under the assumptions of Proposition 6.1, that $|\mathbb{E}_{\boldsymbol{g}_i} T_i| \prec N^{-\frac{\gamma}{4}}$. To see this, we recall (6.8) and (6.11) which together with assumption (6.1) imply that

$$S_i = (a_i - z)G_{ii}^{(i)} - 1 + \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N}}\right) = -\frac{z - \omega_B}{a_i - \omega_B} + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right). \tag{6.25}$$

By the concentration estimate (5.45), we also have

$$\mathbb{E}_{\boldsymbol{g}_i}[S_i] = -\frac{z - \omega_B}{a_i - \omega_B} + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right). \tag{6.26}$$

In addition, by the identity $\widetilde{B}G = I - (A - z)G$, assumption (6.1) and equality (5.38), we have, using the shorthand notation $m_{\boxplus} \equiv m_{\mu_A \boxplus \mu_B}$,

$$\operatorname{tr} G = m_{\boxplus} + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right), \qquad \operatorname{tr}\left(\widetilde{B}G\right) = (z - \omega_B)m_{\boxplus} + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right). \tag{6.27}$$

Substituting (6.26) and assumption (6.1) into (6.24), and using $|T_i|, |S_i| \prec 1$, we obtain

$$\left|\left(1 - \operatorname{tr}\left(\widetilde{B}G\right) + b_i \operatorname{tr} G\right)\mathbb{E}_{\boldsymbol{g}_i}[T_i]\right| \prec N^{-\frac{\gamma}{4}}. \tag{6.28}$$

Using (6.27) and the second equation of (2.13), we have

$$\begin{aligned}\left|\left(1 - \operatorname{tr}\left(\widetilde{B}G\right) + b_i \operatorname{tr} G\right)\right| &= \left|1 + (\omega_B - z + b_i)m_{\boxplus}\right| + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right) \\ &= \left|(-\omega_A + b_i)m_{\boxplus}\right| + \mathrm{O}_{\prec}\left(N^{-\frac{\gamma}{4}}\right).\end{aligned} \tag{6.29}$$

Since $|(-\omega_A + b_i)m_{\boxplus}| \gtrsim 1$ by (3.15), we have from (6.28) that $\mathbb{E}_{\boldsymbol{g}_i}[T_i] \prec N^{-\frac{\gamma}{4}}$. Hence from (5.45), $|T_i| \prec N^{-\frac{\gamma}{4}}$. Then solving (6.23) and (6.24) for $\mathbb{E}_{\boldsymbol{g}_i}[S_i]$, we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{g}_i}[S_i] = &-\frac{\operatorname{tr}\left(\widetilde{B}G\right)}{\operatorname{tr} G}G_{ii}^{(i)} + \left[\frac{\operatorname{tr}\left(\widetilde{B}G\right) - \left(\operatorname{tr}\left(\widetilde{B}G\right)\right)^2}{\operatorname{tr} G} + \operatorname{tr}\left(\widetilde{B}G\widetilde{B}\right)\right]\left(G_{ii}^{(i)} + \mathbb{E}_{\boldsymbol{g}_i}[T_i]\right) \\ &+ \mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right).\end{aligned} \tag{6.30}$$

Averaging over the index i and reorganizing, we get

$$\left|\frac{\operatorname{tr}\left(\widetilde{B}G\right)-\left(\operatorname{tr}\left(\widetilde{B}G\right)\right)^2}{\operatorname{tr}G}+\operatorname{tr}\left(\widetilde{B}G\widetilde{B}\right)\right|=\left|\frac{\frac{1}{N}\sum_{i=1}^N\left(\frac{\operatorname{tr}\left(\widetilde{B}G\right)}{\operatorname{tr}G}G_{ii}^{(i)}+\mathbb{E}_{\boldsymbol{g}_i}\left[S_i\right]\right)+\mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right)}{\frac{1}{N}\sum_{i=1}^N\left(G_{ii}^{(i)}+\mathbb{E}_{\boldsymbol{g}_i}\left[T_i\right]\right)}\right|. \tag{6.31}$$

Now, recalling the concentration of S_i in (5.45) and estimate (6.11), we have

$$\left|\mathbb{E}_{\boldsymbol{g}_i}\left[S_i\right]+(\widetilde{B}^{(i)}G^{(i)})_{ii}\right|\prec\frac{1}{\sqrt{N\eta}}. \tag{6.32}$$

Note that under assumption (6.1), we can use Corollary 5.2 to get (5.44), which together with (6.1) implies that the assumptions in Lemma 4.1 in the case of $i=j=k$ are satisfied. Then, by (4.11) with $i=j=k$ and (6.8), we get

$$G_{ii}^{(i)}=G_{ii}+\mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right),\qquad(\widetilde{B}^{(i)}G^{(i)})_{ii}=(\widetilde{B}G)_{ii}+\mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right), \tag{6.33}$$

for all $i\in[\![1,N]\!]$. Using (6.32) and (6.33) we obtain

$$\left|\frac{1}{N}\sum_{i=1}^N G_{ii}^{(i)}-\operatorname{tr}G\right|\prec\frac{1}{\sqrt{N\eta}},\qquad\left|\frac{1}{N}\sum_{i=1}^N\mathbb{E}_{\boldsymbol{g}_i}\left[S_i\right]+\operatorname{tr}\left(\widetilde{B}G\right)\right|\prec\frac{1}{\sqrt{N\eta}}. \tag{6.34}$$

Substituting (6.34) and assumption (6.1) into the right side of (6.31), and using $|\operatorname{tr}G|\gtrsim 1$ (following from (6.27)) and $|T_i|\prec N^{-\frac{\gamma}{4}}$, we obtain

$$\left|\frac{\operatorname{tr}\left(\widetilde{B}G\right)-\left(\operatorname{tr}\left(\widetilde{B}G\right)\right)^2}{\operatorname{tr}G}+\operatorname{tr}\left(\widetilde{B}G\widetilde{B}\right)\right|\prec\frac{1}{\sqrt{N\eta}}. \tag{6.35}$$

Now, plugging (6.35) back into (6.30) gives

$$\mathbb{E}_{\boldsymbol{g}_i}\left[S_i\right]=-\frac{\operatorname{tr}\left(\widetilde{B}G\right)}{\operatorname{tr}G}G_{ii}^{(i)}+\mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right), \tag{6.36}$$

which together with (6.8) and (6.32) implies that

$$\left(a_i-\omega_B^c\right)G_{ii}^{(i)}=1+\mathrm{O}_{\prec}\left(\frac{1}{\sqrt{N\eta}}\right), \tag{6.37}$$

in light of the definition of $\omega_B^c(z)$ in (2.30). By assumption (6.1) we see that $\omega_B^c(z)=\omega_B(z)+\mathrm{O}_{\prec}(N^{-\frac{\gamma}{4}})$. Hence by (3.15), we also have $\operatorname{Im}\omega_B^c(z)\geq c$ for some positive constant c. Therefore, we get (6.2) from (6.37).

Then (6.36) and (6.2), together with the definition of $\omega_B^c(z)$ in (2.30) and the concentration of S_i in (5.2), imply the estimate of S_i in (6.3).

Substituting (6.36) into (6.24), we strengthen (6.28) to

$$\left|\left(1-\operatorname{tr}\left(\widetilde{B}G\right)+b_i\operatorname{tr}G\right)\mathbb{E}_{\boldsymbol{g}_i}\left[T_i\right]\right|\prec\frac{1}{\sqrt{N\eta}}. \tag{6.38}$$

Using (6.29) again, we obtain from (6.38) that

$$\left|\mathbb{E}_{\boldsymbol{g}_i}\left[T_i\right]\right| \prec \frac{1}{\sqrt{N\eta}},$$

which together with the concentration inequality in (5.45) implies (6.3). Therefore, we complete the proof of Proposition 6.1. □

We conclude this section with the proof of Lemma 6.2.

Proof of Lemma 6.2. We start by invoking the finite-rank perturbation formula (3.3) to get

$$\left|\operatorname{tr} Q_1^{\langle i\rangle} G^{(i)} Q_2^{\langle i\rangle} - \operatorname{tr} Q_1^{\langle i\rangle} G Q_2^{\langle i\rangle}\right| \le \frac{2\|Q_1^{\langle i\rangle} Q_2^{\langle i\rangle}\|}{N\eta}, \qquad Q_1^{\langle i\rangle}, Q_2^{\langle i\rangle} = I \text{ or } \widetilde{B}^{\langle i\rangle}.$$

Hence, it suffices to verify (6.4) and (6.5) with $G^{(i)}$ replaced by G. Recalling from Sect. 4 that $R_i = I - \boldsymbol{r}_i \boldsymbol{r}_i^*$ and using the fact that R_i is a Householder reflection (in fact $\|\boldsymbol{r}_i\|_2^2 = 2$ by construction), we have $\widetilde{B}^{\langle i\rangle} = R_i \widetilde{B} R_i$. Then we write

$$\operatorname{tr}\left(\widetilde{B}^{\langle i\rangle} G\right) = \operatorname{tr}\left(R_i \widetilde{B} R_i G\right) = \operatorname{tr}\left(\widetilde{B} G\right) + d_i, \tag{6.39}$$

with

$$d_i := -\frac{1}{N}\boldsymbol{r}_i^* \widetilde{B} G \boldsymbol{r}_i - \frac{1}{N}\boldsymbol{r}_i^* G \widetilde{B} \boldsymbol{r}_i + \frac{1}{N}(\boldsymbol{r}_i^* \widetilde{B} \boldsymbol{r}_i)(\boldsymbol{r}_i G \boldsymbol{r}_i).$$

Using that $\|G\| \le 1/\eta$, we immediately get the deterministic bound $|d_i| \le C/N\eta$, for some numerical constant C. Together with (6.39) this implies the first estimate in (6.4). The second estimate in (6.4) is obtained in the similar way.

The bounds in (6.5) follow by combining the sharp formulas for $\operatorname{tr}(\widetilde{B}G)$ and $\operatorname{tr}(\widetilde{B}G\widetilde{B})$ from (6.27), (6.35) with the estimates in (6.4).

To prove (6.6), we set $Q^{\langle i\rangle} = \widetilde{B}^{\langle i\rangle}$ or $(\widetilde{B}^{\langle i\rangle})^2$ and note that

$$\left|\mathbb{E}_{\boldsymbol{g}_i}[\operatorname{tr}\left(Q^{\langle i\rangle} G^{(i)}\right)]\right| = \left|\mathbb{E}_{\boldsymbol{g}_i}[\operatorname{tr}\left(Q^{\langle i\rangle}(G^{(i)} - G^{\langle i\rangle})\right)]\right| \le \frac{2\|Q^{\langle i\rangle}\|}{N\eta},$$

where we used that $\boldsymbol{g}_i$ and $G^{\langle i\rangle}$ are independent, and once more (3.3). □

7. Proof of Theorem 2.5: Inequalities (2.20) and (2.22)

In this section, we prove the estimates (2.20) and (2.22) of Theorem 2.5 via a continuity argument. We also prove Theorem 2.6.

First, let us recall the matrix $\mathcal{H}$ and its Green function $\mathcal{G}$ defined in (2.28) and (2.29), these are the natural counterparts of H and G with the roles of A and B as well as the roles of U and U^*interchanged. We can apply a similar partial randomness decomposition to the unitary U^* in $\mathcal{H}$ as we did for U in H in Sect. 4. This means that, for any $i \in [\![1, N]\!]$, there exists an independent pair $(\widehat{\boldsymbol{v}}_i, \mathcal{U}^i)$, uniformly distributed on $\mathcal{S}_{\mathbb{C}}^{N-1}$ and $U(N-1)$, respectively, such that with $\widehat{\boldsymbol{r}}_i := \sqrt{2}(\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\widehat{\theta}_i}\widehat{\boldsymbol{v}}_i)/\|\boldsymbol{e}_i + \mathrm{e}^{-\mathrm{i}\widehat{\theta}_i}\widehat{\boldsymbol{v}}_i\|_2$, we have the decomposition $U^* = -\mathrm{e}^{\mathrm{i}\widehat{\theta}_i}\mathcal{R}_i\,\mathcal{U}^{\langle i\rangle}$, where $\widehat{\theta}_i$ is the argument of the ith coordinate of $\widehat{\boldsymbol{v}}_i$; where $\mathcal{R}_i := (I - \widehat{\boldsymbol{r}}_i\widehat{\boldsymbol{r}}_i^*)$ and $\mathcal{U}^{\langle i\rangle}$ is the unitary matrix with $\boldsymbol{e}_i$ as its ith

column and $\mathcal{U}^i$ as its (i,i)-matrix minor. Analogously to $\boldsymbol{g}_i$ defined in (4.6), we define a Gaussian vector $\widehat{\boldsymbol{g}}_i = (\widehat{g}_{i1}, \dots, \widehat{g}_{iN}) \sim \mathcal{N}_{\mathbb{C}}(0, N^{-1}I)$, to approximate $\mathrm{e}^{-\mathrm{i}\widehat{\theta}_i}\widehat{\boldsymbol{v}}_i$. Setting $\widehat{\boldsymbol{w}}_i := \boldsymbol{e}_i + \widehat{\boldsymbol{g}}_i$ and $\mathcal{W}_i := I - \widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i^*$, we define

$$\mathcal{H}^{(i)} := B + \mathcal{W}_i \mathcal{U}^{\langle i\rangle} A\, (\mathcal{U}^{\langle i\rangle})^* \mathcal{W}_i\,,$$

for all $i \in [\![1, N]\!]$. Calligraphic letters are used to distinguish the decompositions of $\mathcal{H}$ from the decompositions of H.

Next, we introduce the z-dependent random variable

$$\begin{aligned}\Lambda_{\mathrm{d}}(z) := &\max_{i\in[\![1,N]\!]} |G_{ii}^{(i)}(z) - \big(a_i - \omega_B(z)\big)^{-1}| + \max_{i\in[\![1,N]\!]} |G_{ii}(z) - \big(a_i - \omega_B(z)\big)^{-1}| \\ &+ \max_{i\in[\![1,N]\!]} |\mathcal{G}_{ii}^{(i)}(z) - \big(b_i - \omega_A(z)\big)^{-1}| + \max_{i\in[\![1,N]\!]} |\mathcal{G}_{ii}(z) - \big(b_i - \omega_A(z)\big)^{-1}|.\end{aligned} \tag{7.1}$$

Moreover, for any $\delta \in [0,1]$ and $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$, we define the following event

$$\Theta_{\mathrm{d}}(z,\delta) := \{\Lambda_{\mathrm{d}}(z) \le \delta\}. \tag{7.2}$$

The subscript d refers to "diagonal" matrix elements. With the above notation, we have the following lemma.

Lemma 7.1. *Suppose that the assumptions of Theorem 2.5 are satisfied and fix $\gamma > 0$. For any ε with $0 < \varepsilon \le \frac{\gamma}{8}$ and for any $D > 0$ there exists a positive integer $N_2(D,\varepsilon)$ such that the following holds: For any fixed $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ there exists an event $\Omega_{\mathrm{d}}(z) \equiv \Omega_{\mathrm{d}}(z, D, \varepsilon)$ with*

$$\mathbb{P}\big(\Omega_{\mathrm{d}}(z)\big) \ge 1 - N^{-D}, \qquad \forall N \ge N_2(D,\varepsilon)\,,$$

such that if the estimate

$$\mathbb{P}\big(\Theta_{\mathrm{d}}(z, N^{-\frac{\gamma}{4}})\big) \ge 1 - N^{-D}\big(1 + N^5(1-\eta)\big) \tag{7.3}$$

holds for all $D > 0$ and $N \ge N_1(D,\gamma,\varepsilon)$, for some threshold $N_1(D,\gamma,\varepsilon)$, then we also have

$$\Theta_{\mathrm{d}}(z, N^{-\frac{\gamma}{4}}) \cap \Omega_{\mathrm{d}}(z) \subset \Theta_{\mathrm{d}}\Big(z, \frac{N^{\varepsilon}}{\sqrt{N\eta}}\Big) \tag{7.4}$$

for all $N \ge N_3(D,\gamma,\varepsilon) := \max\{N_1(D,\gamma,\varepsilon), N_2(D,\varepsilon)\}$.

Proof. In this proof we fix $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. By the definition of $\prec$ in Definition 1.1, we see that assumption (7.3) implies

$$\big|G_{ii}^{(i)}(z) - (a_i - \omega_B(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}, \qquad \big|G_{ii}(z) - (a_i - \omega_B(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}, \tag{7.5}$$

and

$$\big|\mathcal{G}_{ii}^{(i)}(z) - (b_i - \omega_A(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}, \qquad \big|\mathcal{G}_{ii}(z) - (b_i - \omega_A(z))^{-1}\big| \prec N^{-\frac{\gamma}{4}}. \tag{7.6}$$

Hence, we can use Corollary 5.2 to get (5.44). Together with the boundedness of $G_{ii}^{(i)}$ and G_{ii} (*c.f.* (7.5) and (3.15)) this implies that the assumptions in (4.10) of Lemma 4.1

are satisfied when $i = j = k$. Thus (4.11) holds when $i = j = k$. Hence, invoking, (7.5) and Proposition 6.1, we get

$$\left|G^{(i)}_{ii}(z) - (a_i - \omega^c_B(z))^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}, \qquad \left|G_{ii}(z) - (a_i - \omega^c_B(z))^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{7.7}$$

Switching the roles of A and B as well as U and U^*, and further using (2.31), we also get

$$\left|\mathcal{G}^{(i)}_{ii}(z) - (b_i - \omega^c_A(z))^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}, \qquad \left|\mathcal{G}_{ii}(z) - (b_i - \omega^c_A(z))^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}, \tag{7.8}$$

under (7.6).

Now, we state the conclusions (7.7) and (7.8) in a more explicit quantitative form assuming (7.3) which is a quantitative form of (7.5)–(7.6). Namely, we show that the inequalities

$$\begin{aligned}
&\left|G^{(i)}_{ii}(z) - (a_i - \omega^c_B(z))^{-1}\right| \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad \left|G_{ii}(z) - (a_i - \omega^c_B(z))^{-1}\right| \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \\
&\left|\mathcal{G}^{(i)}_{ii}(z) - (b_i - \omega^c_A(z))^{-1}\right| \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad \left|\mathcal{G}_{ii}(z) - (b_i - \omega^c_A(z))^{-1}\right| \le \frac{N^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}.
\end{aligned} \tag{7.9}$$

hold on the event $\Theta_{\mathrm{d}}(z, N^{-\frac{\gamma}{4}}) \cap \Omega_{\mathrm{d}}(z)$, when $N \ge N_3(D, \gamma, \varepsilon)$. Here $\Omega_{\mathrm{d}}(z)$ is an event determined as the intersection of the "typical" events in all the concentration estimates in Sects. 4–6.

To see this more precisely, we go back to the proofs in these sections. The concentration estimates always involved quantities of the form $\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* Q \boldsymbol{x}]$ with $\boldsymbol{x} = \boldsymbol{g}_i, \boldsymbol{e}_i$ and some explicit matrix Q that is independent of $\boldsymbol{g}_i$ but often z-dependent. The total number of such estimates was linear in N. Thus, according to Lemma 3.1, for any (small) $\varepsilon > 0$ and (large) $D > 0$, there exists an event $\Omega_{\mathrm{d}}(z, D, \varepsilon)$ with

$$\mathbb{P}\big(\Omega_{\mathrm{d}}(z, D, \varepsilon)\big) \ge 1 - N^{-D} \tag{7.10}$$

such that all estimates of the form

$$|\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* Q \boldsymbol{e}_i]| \le \frac{N^{\frac{\varepsilon}{4}}}{\sqrt{N}} \|Q\boldsymbol{e}_i\|_2, \qquad |\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* Q \boldsymbol{g}_i]| \le \frac{N^{\frac{\varepsilon}{4}}}{N} \|Q\|_2 \tag{7.11}$$

in Sects. 4–6 hold on $\Omega_{\mathrm{d}}(z, D, \varepsilon)$ for all $N \ge N_2(D, \varepsilon)$. In addition, the threshold $N_2(D, \varepsilon)$ is independent of the spectral parameter z.

We now follow the proofs in Sects. 4–6 to the letter but we use (7.10), (7.11) and (7.3) instead of the $\prec$ relation. Instead of (7.7) and (7.8), we find that the analogous but more quantitative bounds (7.9) hold on the intersection of the events $\Theta_{\mathrm{d}}(z, N^{-\frac{\gamma}{4}})$ and $\Omega_{\mathrm{d}}(z, D, \varepsilon)$.

It remains to show that on the event $\Theta_{\mathrm{d}}(z, N^{-\frac{\gamma}{4}}) \cap \Omega_{\mathrm{d}}(z)$,

$$|\omega^c_A(z) - \omega_A(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |\omega^c_B(z) - \omega_B(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}} \tag{7.12}$$

hold when $N \ge N_3(D, \gamma, \varepsilon)$.

To this end, we use the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A,\omega_B,z)=0$ as formulated in Lemma 3.3. By the definition of the approximate subordination functions $\omega_A^c(z)$ and $\omega_B^c(z)$ in (2.30), by the identity (2.32) and by taking the average over the index i in the estimates in (7.9), we get the system of equations

$$\begin{aligned} m_H(z) &= m_A(\omega_B^c(z)) + r_A(z), \\ m_H(z) &= m_B(\omega_A^c(z)) + r_B(z), \\ \omega_A^c(z) + \omega_B^c(z) &= z - \frac{1}{m_H(z)}, \end{aligned} \tag{7.13}$$

where the error terms r_A and r_B satisfy

$$|r_A(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |r_B(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}},$$

on the event $\Theta_{\mathrm{d}}(z,N^{-\frac{\gamma}{4}})\cap\Omega_{\mathrm{d}}(z)$ when $N\ge N_3(D,\gamma,\varepsilon)$. Using the definition of $\Theta_{\mathrm{d}}(z,\delta)$ in (7.2), (7.9) and the fact that $z\in\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}},1)$, so $\omega_A(z)$ and $\omega_B(z)$ are well separated from the real axis, we have

$$|\omega_A^c(z)-\omega_A(z)| \le CN^{-\frac{\gamma}{4}}, \qquad |\omega_B^c(z)-\omega_B(z)| \le CN^{-\frac{\gamma}{4}}. \tag{7.14}$$

on the event $\Theta_{\mathrm{d}}(z,N^{-\frac{\gamma}{4}})\cap\Omega_{\mathrm{d}}(z)$ when $N\ge N_3(D,\gamma,\varepsilon)$. Hence, plugging the third equation of (7.13) into the first two and using (3.15) together with (7.14), we get

$$\Phi_{\mu_A,\mu_B}(\omega_A^c(z),\omega_B^c(z),z)=\widetilde{r}(z),$$

where $\widetilde{r}(z)=(\widetilde{r}_A(z),\widetilde{r}_B(z))$ with

$$|\widetilde{r}_A(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}}, \qquad |\widetilde{r}_B(z)| \le \frac{CN^{\frac{\varepsilon}{2}}}{\sqrt{N\eta}} \tag{7.15}$$

on the event $\Theta_{\mathrm{d}}(z,N^{-\frac{\gamma}{4}})\cap\Omega_{\mathrm{d}}(z)$ when $N\ge N_3(D,\gamma,\varepsilon)$. Therefore, by Lemma 3.3, we get (7.12). Hence, we completed the proof of Lemma 7.1. □

Given Lemma 7.1, we next prove Theorem 2.5 via a continuity argument similarly to [19].

Proof of (2.20) *of Theorem 2.5.* Using Theorem 1.2 (i) of [26] together with Lemma C.1 of [26], we see that for $\eta=1$, we have

$$\begin{aligned} \max_{i\in[\![1,N]\!]} \left|G_{ii}(z)-(a_i-\omega_B(z))^{-1}\right| &\prec N^{-\frac{\gamma}{2}}, \\ \max_{i\in[\![1,N]\!]} \left|\mathcal{G}_{ii}(z)-(b_i-\omega_A(z))^{-1}\right| &\prec N^{-\frac{\gamma}{2}} \end{aligned} \tag{7.16}$$

if $0<\gamma\le 1/7$ (say). In addition, owing to the estimate $\|G\|\le 1/\eta$, assumption (4.10) obviously holds for $\eta=1$. Hence, by Lemma 4.1 in the case of $i=j=k$ and its analogue for $\mathcal{G}_{ii}^{(i)}$, we have

$$\begin{aligned} \max_{i\in[\![1,N]\!]} \left|G_{ii}^{(i)}(z)-(a_i-\omega_B(z))^{-1}\right| &\prec N^{-\frac{\gamma}{2}}, \\ \max_{i\in[\![1,N]\!]} \left|\mathcal{G}_{ii}^{(i)}(z)-(b_i-\omega_A(z))^{-1}\right| &\prec N^{-\frac{\gamma}{2}}. \end{aligned} \tag{7.17}$$

Hence, for any $E \in \mathcal{I}$ and $D > 0$,

$$\mathbb{P}\big(\Theta_{\mathrm{d}}(E + \mathrm{i}, N^{-\frac{3\gamma}{8}})\big) \geq 1 - N^{-D}, \tag{7.18}$$

holds for all $N \geq N_0(D, \gamma)$ with some $N_0(D, \gamma) > 0$. In the sequel we will apply Lemma 7.1 with the choice

$$N_1(D, \gamma, \varepsilon) := \max\big\{N_0(D, \gamma), N_2(D, \varepsilon)\big\};$$

in particular we have $N_3(D, \gamma, \varepsilon) = N_1(D, \gamma, \varepsilon)$.

Next, we define the lattice

$$\widehat{\mathcal{S}}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1) := \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1) \cap N^{-5}\{\mathbb{Z} \times \mathrm{i}\mathbb{Z}\}. \tag{7.19}$$

Thanks to the Lipschitz continuity of the Green function, *i.e.* $\|G(z) - G(z')\| \leq N^2|z - z'|$ for any $z, z' \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$, and of the subordination functions (see (3.9)), it suffices to show (2.20) on the lattice $\widehat{\mathcal{S}}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. We now fix $E \in \mathcal{I} \cap N^{-5}\mathbb{Z}$ and decrease η from $\eta = 1$ down to $N^{-1+\gamma}$ in steps of size N^{-5}. Recall the events $\Theta_{\mathrm{d}}(z, \delta)$ and $\Omega_{\mathrm{d}}(z)$ in Lemma 7.1, and choose the same $\varepsilon < \frac{\gamma}{8}$ in $\Omega_{\mathrm{d}}(z, D, \varepsilon)$ for all z. For simplicity, we omit the real part E from the notation and rewrite

$$\Theta_{\mathrm{d}}(\eta, \delta) := \Theta_{\mathrm{d}}(E + \mathrm{i}\eta, \delta), \qquad \Omega_{\mathrm{d}}(\eta) := \Omega_{\mathrm{d}}(E + \mathrm{i}\eta).$$

Our aim is to show that for any $\eta \in [\eta_{\mathrm{m}}, 1]$,

$$\Theta_{\mathrm{d}}\big(\eta, N^{-\frac{3\gamma}{8}}\big) \cap \Omega_{\mathrm{d}}(\eta - N^{-5}) \subset \Theta_{\mathrm{d}}\big(\eta - N^{-5}, N^{-\frac{3\gamma}{8}}\big). \tag{7.20}$$

To see (7.20), we first notice that by the Lipschitz continuity of the Green function and of the subordination functions $\omega_A(z)$ and $\omega_B(z)$ (see (3.9)), we have

$$\Theta_{\mathrm{d}}\big(\eta, N^{-\frac{3\gamma}{8}}\big) \subset \Theta_{\mathrm{d}}\big(\eta - N^{-5}, N^{-\frac{3\gamma}{8}} + CN^{-3}\big) \subset \Theta_{\mathrm{d}}\big(\eta - N^{-5}, N^{-\frac{\gamma}{4}}\big), \tag{7.21}$$

where the last step is obtained by choosing $\gamma > 0$ sufficiently small. Now, we start from (7.18). By (7.21), we get

$$\mathbb{P}\big(\Theta_{\mathrm{d}}\big(1 - N^{-5}, N^{-\frac{\gamma}{4}}\big)\big) \geq \mathbb{P}\big(\Theta_{\mathrm{d}}\big(1, N^{-\frac{3\gamma}{8}}\big)\big) \geq 1 - N^{-D}.$$

Hence, we can use Lemma 7.1 to get

$$\begin{aligned}\Theta_{\mathrm{d}}\big(1 - N^{-5}, N^{-\frac{\gamma}{4}}\big) \cap \Omega_{\mathrm{d}}(1 - N^{-5}) &\subset \Theta_{\mathrm{d}}\left(1 - N^{-5}, \frac{N^{\varepsilon}}{\sqrt{N(1 - N^{-5})}}\right) \\ &\subset \Theta_{\mathrm{d}}\left(1 - N^{-5}, N^{-\frac{3\gamma}{8}}\right),\end{aligned} \tag{7.22}$$

which together with (7.21) implies (7.20) with $\eta = 1$. Now, replacing 1 by $1 - N^{-5}$, we get from (7.22), (7.18) and the fact $\mathbb{P}(\Omega_{\mathrm{d}}(1 - N^{-5})) \geq 1 - N^{-D}$ for $N \geq N_2(D, \varepsilon)$ that

$$\mathbb{P}\big(\Theta_{\mathrm{d}}\big(1 - N^{-5}, N^{-\frac{3\gamma}{8}}\big)\big) \geq 1 - 2N^{-D} \tag{7.23}$$

holds for all $N \geq N_3(D, \gamma, \varepsilon)$. Now, using (7.23) instead of (7.18), we get (7.20) for $\eta = 1 - N^{-5}$. Iterating this argument, we obtain for any $\eta \in [\eta_{\mathrm{m}}, 1] \cap N^{-5}\mathbb{Z}$ that

$$\Theta_{\mathrm{d}}\big(1, N^{-\frac{3\gamma}{8}}\big) \cap \Omega_{\mathrm{d}}(1 - N^{-5}) \cap \cdots \cap \Omega_{\mathrm{d}}(\eta) \subset \Theta_{\mathrm{d}}\big(\eta, N^{-\frac{3\gamma}{8}}\big).$$

Hence, we have

$$\mathbb{P}\big(\Theta_{\mathrm{d}}\big(\eta, N^{-\frac{3\gamma}{8}}\big)\big) \geq 1 - N^{-D}\big(1 + N^5(1-\eta)\big)$$

for all $N \geq N_3(D, \gamma, \varepsilon)$, which further implies

$$\mathbb{P}\big(\Theta_{\mathrm{d}}\big(\eta - N^{-5}, N^{-\frac{\gamma}{4}}\big)\big) \geq 1 - N^{-D}\big(1 + N^5(1-\eta)\big)$$

for all $N \geq N_3(D, \gamma, \varepsilon)$, by using (7.21). Then, using Lemma 7.1 again, we obtain

$$\mathbb{P}\Big(\Theta_{\mathrm{d}}\Big(\eta, \frac{N^{\varepsilon}}{\sqrt{N\eta}}\Big)\Big) \geq 1 - N^{-D}\big(2 + N^5(1-\eta)\big) \tag{7.24}$$

uniformly for all $\eta \in [\eta_{\mathrm{m}}, 1] \cap N^{-5}\mathbb{Z}$, when $N \geq N_3(D, \gamma, \varepsilon)$. Finally, by continuity, we can extend the bounds from z in the discrete lattice to the entire domain $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. We then get

$$\max_{i\in[\![1,N]\!]}\Big|G_{ii}(z) - \frac{1}{a_i - \omega_B(z)}\Big| \prec \frac{1}{\sqrt{N\eta}}, \qquad \max_{i\in[\![1,N]\!]}\Big|G^{(i)}_{ii}(z) - \frac{1}{a_i - \omega_B(z)}\Big| \prec \frac{1}{\sqrt{N\eta}}, \tag{7.25}$$

uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$, where we used the definitions of $\Theta_{\mathrm{d}}(z, \delta)$ in (7.2) and of $\prec$ in Definition 1.1. This concludes the proof of (2.20). □

Having established (2.20), (2.22) of Theorem 2.5 and Theorem 2.6 are direct consequences.

Proof of (2.22) *of Theorem 2.5.* It suffices to note that (2.22) is a direct consequence of (2.20) and the facts $m_H(z) = N^{-1}\sum_{i=1}^N G_{ii}(z)$ and $m_{A\boxplus B}(z) = N^{-1}\sum_{i=1}^N (a_i - \omega_B)^{-1}$. □

Proof of Theorem 2.6. Using the spectral decomposition of the Green function G, we have

$$\max_{j\in[\![1,N]\!]} \operatorname{Im} G_{jj}(z) = \max_{j\in[\![1,N]\!]} \sum_{i=1}^N \frac{|u_{ij}|^2\eta}{|\lambda_i - E|^2 + \eta^2} = \sum_{i=1}^N \frac{||\boldsymbol{u}_i||_\infty^2\eta}{|\lambda_i - E|^2 + \eta^2}, \qquad z \in \mathbb{C}^+. \tag{7.26}$$

Fix a small $\gamma > 0$. For any $\lambda_i \in \mathcal{I}$, we set $E = \lambda_i$ on the right side of (7.26) and use (2.20) to bound the left side of it with $z = \lambda_i + \mathrm{i}\eta$, $\eta = N^{-1+\gamma}$. Then we obtain

$$||\boldsymbol{u}_i||_\infty^2 \prec \eta = N^{-1+\gamma}.$$

Since $\gamma > 0$ is arbitrarily small, we get (2.26). This completes the proof of Theorem 2.6. □

8. Proof of Theorem 2.5: Inequalities (2.21)

In this section, we prove (2.21) of Theorem 2.5. Note that, from (7.25) in the proof of (2.20) in Theorem 2.5, we know that the following estimates hold uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$,

$$\left|G_{ii}^{(i)}(z) - \big(a_i - \omega_B(z)\big)^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}, \quad \left|G_{ii}(z) - \big(a_i - \omega_B(z)\big)^{-1}\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.1}$$

Taking (8.1) as an input, we follow the discussion in Sects. 5–7 to prove the estimate (2.21) with the following modifications. We introduce the quantities

$$T_{i,j}(z) := \boldsymbol{g}_i^* G^{(i)}(z)\boldsymbol{e}_j, \qquad S_{i,j}(z) := \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)}(z)\boldsymbol{e}_j, \qquad z \in \mathbb{C}^+. \tag{8.2}$$

that generalize $T_i(z)$ and $S_i(z)$ defined in (5.1). In particular, $T_i(z) \equiv T_{i,i}(z)$ and $S_i(z) \equiv S_{i,i}(z)$, but we henceforth implicitly assume that $i \neq j$. (We use a comma in the subscripts of $T_{i,j}$, $S_{i,j}$ since they are not the entries of some matrix.) We often abbreviate $T_{i,j} \equiv T_{i,j}(z)$ and $S_{i,j} \equiv S_{i,j}(z)$.

We first establish the concentration estimates for $G_{ij}^{(i)}$ (see Lemma 8.1), and $T_{i,j}$ and $S_{i,j}$; see Lemma 8.2. In Proposition 8.3 we then derive self-consistent equations for $\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]$ and $\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big]$ that will show, together with concentration estimates, that $|G_{ij}^{(i)}|, |T_{i,j}|, |S_{i,j}| \prec \frac{1}{\sqrt{N\eta}}$, provided that $|G_{ij}^{(i)}| \prec 1$. We then close the argument via continuity.

We start with the analogue of Lemma 5.1 for the off-diagonal entries of $G^{(i)}$.

Lemma 8.1. *Suppose that the assumptions of Theorem 2.5 are satisfied and let $\gamma > 0$. Fix $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ and assume that*

$$|G_{ij}^{(i)}(z)| \prec 1, \tag{8.3}$$

for all $i, j \in [\![1, N]\!]$, $i \neq j$. Then

$$\left|\mathbb{E}_{\boldsymbol{g}_i}[G_{ij}^{(i)}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}, \tag{8.4}$$

for all $i, j \in [\![1, N]\!]$, $i \neq j$.

Proof of Lemma 8.1. Recall $H^{[i]}$ and $H^{\{i\}}$ defined in (5.12) and (5.6), as well as their Green functions $G^{[i]}$ and $G^{\{i\}}$. In the proof of Lemma 5.1 we derived the identity

$$G_{ij}^{(i)}(z) = \frac{G_{ij}^{\{i\}}(z) - (b_i + \omega_B(z) - z)\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)]\, G_{ij}^{(i)}(z)}{1 + (b_i + \omega_B(z) - z)\mathbb{E}_{\boldsymbol{g}_i}[G_{ii}^{\{i\}}(z)]}; \tag{8.5}$$

see (5.8). With (8.1), we see that assumption (5.3) is satisfied. Hence, we can use all the conclusions in the proof of Lemma 5.1. Therefore, according to (5.11), (5.10) and assumption (8.3), it suffices to show the concentration estimate $|\mathbb{E}_{\boldsymbol{g}_i}[G_{ij}^{\{i\}}(z)]| \prec \frac{1}{\sqrt{N\eta}}$. To this end we expand $G^{\{i\}}$ around $G^{[i]}$. Recall from (5.19) that

$$G_{ij}^{\{i\}} = G_{i,j}^{[i]} + \frac{\Psi_{i,j}}{1 + \Xi_i},$$

where Ξ_i is defined in (5.18) and $\Psi_{i,j}$ is defined in (5.20). Recalling statements (i) and (ii) in (5.22), it suffices to establish that $|\mathbb{E}_{\boldsymbol{g}_i}[\Psi_{i,j}]| \prec \frac{1}{\sqrt{N\eta}}$. Note that $\Psi_{i,j}$ contains the terms listed in (5.34), (5.35) and (5.42), as well as the terms

$$\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j\,, \qquad \boldsymbol{g}_i^* G^{[i]} \boldsymbol{e}_j\,, \qquad \boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j. \tag{8.6}$$

Since $\boldsymbol{e}_i$ is an eigenvector of $\widetilde{B}^{\langle i\rangle}$ and of $H^{[i]}$, we have $\boldsymbol{e}_i^* \widetilde{B}^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j = \delta_{ij} b_i G^{[i]}_{ii}$. Moreover, using Lemma 3.1 with $Q^{\langle i\rangle} = I$ or $\widetilde{B}^{\langle i\rangle}$, we have

$$\left|\boldsymbol{g}_i^* Q^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j\right| \prec \frac{1}{\sqrt{N}} \|Q^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j\|_2 \prec \frac{1}{\sqrt{N}} \|G^{[i]} \boldsymbol{e}_j\|_2 = \left(\frac{\operatorname{Im} G^{[i]}_{jj}}{N\eta}\right)^{\frac{1}{2}}. \tag{8.7}$$

To control $G^{[i]}_{jj}$, we recall from (5.12) that the matrix $H^{[i]}$ is block-diagonal and we thus have, for $j \neq i$

$$G^{[i]}_{jj} = (A^i + U^i B^i (U^i)^* - z I_{N-1})^{-1}_{\ell\ell}\,, \qquad \ell := j\mathbf{1}(j < i) + (j-1)\mathbf{1}(j > i)\,,$$

where A^i and B^i the are (i,i)-matrix minors of A and B respectively (obtained by removing the ith column and ith row) and $U^i \in U(N-1)$ is the (i,i)-matrix minor of $U^{\langle i\rangle}$ which is Haar distributed as seen at the beginning of Sect. 4. Note that the matrix $A^i + U^i B^i (U^i)^*$ satisfies the assumptions of Theorem 2.5. We thus have the estimate

$$\max_{i\neq j} \left|G^{[i]}_{jj}(z)\right| \prec 1\,, \qquad z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1). \tag{8.8}$$

Plugging this bound into (8.7) we get $|\boldsymbol{g}_i^* Q^{\langle i\rangle} G^{[i]} \boldsymbol{e}_j| \prec \frac{1}{\sqrt{N\eta}}$. The remaining part of the proof is nearly the same as the one of Lemma 5.1. We omit the details. □

We have the following analogue of Corollary 5.2.

Lemma 8.2. *Suppose that the assumptions of Theorem 2.5 are satisfied and let $\gamma > 0$. For all $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$, we have the bounds*

$$\max_{i\neq j} |T_{i,j}(z)| \prec 1\,, \qquad \max_{i\neq j} |S_{i,j}(z)| \prec 1 \tag{8.9}$$

and the concentration estimates

$$\max_{i\neq j} \left|\mathbb{E}_{\boldsymbol{g}_i}[T_{i,j}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}\,, \qquad \max_{i\neq j} \left|\mathbb{E}_{\boldsymbol{g}_i}[S_{i,j}(z)]\right| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.10}$$

Proof. With the estimates in (8.1) and (8.8), the proof is analogous to that of Corollary 5.2. Here we get the conclusions for all $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ at once, since we use the uniform estimate (8.1) instead of assumption (5.43) for one fixed z. We omit the details.

Finally, we have the following counterpart to Proposition 6.1.

Proposition 8.3. *Fix $z = E + \mathrm{i}\eta \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. Under the assumptions of Lemma 8.1, we have*

$$\max_{i\neq j} |G^{(i)}_{ij}(z)| \prec \frac{1}{\sqrt{N\eta}} \tag{8.11}$$

and

$$\max_{i\neq j} |T_{i,j}(z)| \prec \frac{1}{\sqrt{N\eta}}\,, \qquad \max_{i\neq j} |S_{i,j}(z)| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.12}$$

Proof. The proof is similar to that of Proposition 6.1. Having established the concentration inequalities in (8.4), it suffices to estimate $\mathbb{E}_{\boldsymbol{g}_i}\big[G^{(i)}_{ij}\big]$ to prove (8.11). We then start with

$$(a_i - z)G^{(i)}_{ij} = -(\widetilde{B}^{(i)}G^{(i)})_{ij} + \delta_{ij}. \tag{8.13}$$

Choosing henceforth $i \neq j$, mimicking the reasoning from (6.9) to (6.11) and using (8.9), we arrive at

$$(\widetilde{B}^{(i)}G^{(i)})_{ij} = -\boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}G^{(i)}\boldsymbol{e}_j + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N}}\Big) = -S_{i,j} + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N}}\Big). \tag{8.14}$$

Then, instead of (6.17), we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}[S_{i,j}] &= \mathbb{E}_{\boldsymbol{g}_i}\Big[\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\big)\big(S_{i,j} - b_i\,T_{i,j}\big)\Big] \\
&\quad + \mathbb{E}_{\boldsymbol{g}_i}\Big[\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\widetilde{B}^{\langle i\rangle}\big)\big(G^{(i)}_{ij} + T_{i,j}\big)\Big] \\
&\quad - \mathbb{E}_{\boldsymbol{g}_i}\Big[\mathrm{tr}\,\big(\widetilde{B}^{\langle i\rangle}G^{(i)}\big)\,\big(\boldsymbol{e}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i + \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{e}_i + \boldsymbol{g}_i^*\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\big)\big(G^{(i)}_{ij} + T_{i,j}\big)\Big] \\
&\quad - \frac{1}{N}\mathbb{E}_{\boldsymbol{g}_i}\Big[\Big(b_i^2G^{(i)}_{ii} + \boldsymbol{g}_i^*\big(\widetilde{B}^{\langle i\rangle}\big)^2G^{(i)}\boldsymbol{e}_i + \boldsymbol{e}_i^*(\widetilde{B}^{\langle i\rangle})^2G^{(i)}\boldsymbol{g}_i + \boldsymbol{g}_i^*\big(\widetilde{B}^{\langle i\rangle}\big)^2G^{(i)}\boldsymbol{g}_i\Big) \\
&\qquad\qquad \times \big(G^{(i)}_{ij} + T_{i,j}\big)\Big],
\end{aligned} \tag{8.15}$$

where we directly used the definitions in (8.2). Then, similarly to (6.23), using the concentration estimates in Lemma 8.1 and in Lemma 8.2, as well as the Gaussian concentration estimates in (6.10), the bound (6.18) and Lemma 6.2 for tracial quantities, we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big] - \mathrm{tr}\,\big(\widetilde{B}G\widetilde{B}\big)\big(G^{(i)}_{ij} + \mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big) \\
&\quad = \mathrm{tr}\,\big(\widetilde{B}G\big)\,\big(\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j} - b_i\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big]\big) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big).
\end{aligned} \tag{8.16}$$

Analogously, we also have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big] - \mathrm{tr}\,\big(\widetilde{B}G\big)\,\big(G^{(i)}_{ij} + \mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big) \\
&\quad = \mathrm{tr}\,\big(G\big)\,\big(\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big] - b_i\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big) + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big).
\end{aligned} \tag{8.17}$$

Solving $\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big]$ from (8.16) and (8.17), we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big] &= -\frac{\mathrm{tr}\,\big(\widetilde{B}G\big)}{\mathrm{tr}\,G}G^{(i)}_{ij} + \Big[\frac{\mathrm{tr}\,\big(\widetilde{B}G\big) - \big(\mathrm{tr}\,\big(\widetilde{B}G\big)\big)^2}{\mathrm{tr}\,G} + \mathrm{tr}\,\big(\widetilde{B}G\widetilde{B}\big)\Big]\big(G^{(i)}_{ij} + \mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big) \\
&\quad + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big).
\end{aligned}$$

Using (6.35), the assumption $|G^{(i)}_{ij}| \prec 1$ and the bound $|T_{i,j}| \prec 1$ of (8.9), we have

$$\mathbb{E}_{\boldsymbol{g}_i}\big[S_{i,j}\big] = -\frac{\mathrm{tr}\,\big(\widetilde{B}G\big)}{\mathrm{tr}\,G}G^{(i)}_{ij} + \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big), \tag{8.18}$$

which together with (8.13), (8.14), the concentration estimate (8.10) implies that

$$(a_i - \omega_B^c) G_{ij}^{(i)} = \mathrm{O}_{\prec}\Big(\frac{1}{\sqrt{N\eta}}\Big). \tag{8.19}$$

This proves the estimate in (8.11).

Next, we bound $S_{i,j}$. Starting from (8.18) we directly get the second estimates in (8.12) from the Green function bound (8.11) and the concentration estimate (8.10).

It remains to estimate $T_{i,j}$. Plugging the bound on G_{ij} in (8.11) and the bound on $S_{i,j}$ in (8.12) into the equation (8.17), we obtain

$$\big|\big(1 - \mathrm{tr}\,\big(\widetilde{B}G\big) + b_i \mathrm{tr}\, G\big)\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.20}$$

Invoking the estimate (6.29) we get $\big|\mathbb{E}_{\boldsymbol{g}_i}\big[T_{i,j}\big]\big| \prec \frac{1}{\sqrt{N\eta}}$. Then the first estimate in (8.12) follows from the concentration estimate for $T_{i,j}$ in (8.10). This completes the proof. □

Having established Lemma 8.1 and Proposition 8.3, we next prove (2.21) of Theorem 2.5 via a continuity argument similar to the proof of (2.20).

Proof of (2.21) *of Theorem 2.5.* Fixing any $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ and using Proposition 8.3, under the assumption

$$\max_{i \neq j} |G_{ij}^{(i)}(z)| \prec 1\,, \tag{8.21}$$

we have

$$\max_{i \neq j} |G_{ij}^{(i)}(z)| \prec \frac{1}{\sqrt{N\eta}}\,, \qquad \max_{i \neq j} |T_{i,j}(z)| \prec \frac{1}{\sqrt{N\eta}}\,, \qquad \max_{i \neq j} |S_{i,j}(z)| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.22}$$

Then, by (8.22) and (8.1), we can use Lemma 4.1 to get

$$|G_{ij}(z)| \prec \frac{1}{\sqrt{N\eta}}. \tag{8.23}$$

Hence, in principle, it suffices to conduct a continuity argument from $\eta = 1$ to $\eta = \eta_{\mathrm{m}}$ (similar to the proof of (2.20) of Theorem 2.5) to show that the bound (8.21) holds uniformly for $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$. However, in order to show that (8.23) also holds uniformly for $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ quantitatively, we monitor G_{ij} in the continuity argument as well. To this end, we introduce the z-dependent random variable

$$\Lambda_{\mathrm{o}} \equiv \Lambda_{\mathrm{o}}(z) := \max_{i \neq j} |G_{ij}^{(i)}(z)| + \max_{i \neq j} |G_{ij}(z)|\,,$$

and, for any $\delta \in [0, 1]$ and $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$, we define the event

$$\Theta_{\mathrm{o}}(z, \delta) := \{\Lambda_{\mathrm{o}}(z) \leq \delta\}\,;$$

c.f. (7.1) and (7.2). The subscript o refers to "off-diagonal".

We will mimic the proof of (2.20). Analogously, using Lemma 4.1 and Proposition 8.3, one shows that there exists an event $\Omega_{\mathrm{o}}(z) \equiv \Omega_{\mathrm{o}}(z, D, \varepsilon)$ such that the conclusions in Lemma 7.1 still hold when we replace $\Theta_{\mathrm{d}}(z, \delta)$ by $\Theta_{\mathrm{o}}(z, \delta)$, $\Omega_{\mathrm{d}}(z)$ by $\Omega_{\mathrm{o}}(z)$

and $N^{-\frac{\gamma}{4}}$ by 1. We also set $\delta = 1$ in this proof. This is a quantitative description of the derivation of the first bound in (8.22) and (8.23) from (8.21). The main difference is that here $\Omega_{\mathrm{o}}(z)$ is the event defined as the intersection of the "typical" events in all the concentration estimates in Sects. 4–6, in the proofs of Lemma 8.1 and Proposition 8.3, and the event on which the following bounds hold

$$\left|G_{ii}^{(i)}(z) - \big(a_i - \omega_B(z)\big)^{-1}\right| \le \frac{N^{\varepsilon}}{\sqrt{N\eta}}\,,$$
$$\left|G_{ii}(z) - \big(a_i - \omega_B(z)\big)^{-1}\right| \le \frac{N^{\varepsilon}}{\sqrt{N\eta}}\,, \qquad \max_{i\neq j}\left|G_{jj}^{[i]}(z)\right| \le N^{\varepsilon}. \tag{8.24}$$

Note that, by (8.1) and (8.8), we know that (8.24) holds with high probability uniformly on $\mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$.

With the analogue of Lemma 7.1 for $\Theta_{\mathrm{o}}(z, \delta = 1)$ and $\Omega_{\mathrm{o}}(z)$, we conduct a continuity argument similar to the one in the proof of (2.20). Again, by Lipschitz continuity of the Green function it suffices to show estimate (2.21) on the lattice $\widehat{\mathcal{S}}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$ defined in (7.19). We fix $E \in \mathcal{I} \cap N^{-5}\mathbb{Z}$, write $z = E + \mathrm{i}\eta$ and decrease η from $\eta = 1$ down to $N^{-1+\gamma}$ in steps of size N^{-5}. The initial estimate for $\eta = 1$, *i.e.* $\Lambda_{\mathrm{o}}(E + \mathrm{i}) \le 1$ follows directly from the trivial fact $\|G^{(i)}(z)\|, \|G(z)\| \le 1/\eta$. Then one can show step by step that for any $\eta \in [\eta_{\mathrm{m}}, 1]$, say,

$$\Theta_{\mathrm{o}}\big(\eta, 1\big) \cap \Omega_{\mathrm{o}}(\eta - N^{-5}) \subset \Theta_{\mathrm{o}}\Big(\eta - N^{-5}, 1\Big)\,, \tag{8.25}$$

which is the analogue of (7.20). The remaining proof is nearly the same as the counterpart in the proof of (2.20). We thus omit the details. □

Acknowledgements. Open access funding provided by Institute of Science and Technology (IST Austria).

Appendix A. Orthogonal Case

In this appendix, we show that Theorem 2.5 and Theorem 2.6 also hold in the orthogonal setup where U is Haar distributed on the orthogonal group $O(N)$. From the proof of Theorem 2.6, we see that it is implied by Theorem 2.5. Hence, it suffices to discuss the latter. We outline the necessary changes in the discussion of Sects. 4–8 to adapt our proof to the orthogonal case. We mainly show the modification for the proof of (2.20) in detail, and (2.21) will be discussed briefly at the end.

First, we modify some notation. We start with the decomposition for the Haar measure on the orthogonal group analogous to (2.37). For all $i \in [\![1, N]\!]$, according to [28], there exist a random vector $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iN})$, uniformly distributed on the real unit $(N-1)$-sphere $\mathcal{S}_{\mathbb{R}}^{N-1} := \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^*\boldsymbol{x} = 1\}$, and a Haar distributed orthogonal matrix $U^i \in O(N-1)$, which is independent of $\boldsymbol{v}_i$, such that one has the decomposition

$$U = -\operatorname{sgn}(v_{ii})(I - \boldsymbol{r}_i\boldsymbol{r}_i^*)U^{\langle i\rangle} := -\operatorname{sgn}(v_{ii})R_i\,U^{\langle i\rangle}\,,$$

where

$$\boldsymbol{r}_i := \sqrt{2}\frac{\boldsymbol{e}_i + \operatorname{sgn}(v_{ii})\boldsymbol{v}_i}{\|\boldsymbol{e}_i + \operatorname{sgn}(v_{ii})\boldsymbol{v}_i\|_2}\,, \quad R_i := I - \boldsymbol{r}_i\boldsymbol{r}_i^*\,, \tag{A.1}$$

and $U^{\langle i\rangle}$ is the orthogonal matrix with $\boldsymbol{e}_i$ as its ith column and U^i as its (i,i)-matrix minor. Moreover, there is a real Gaussian vector $\boldsymbol{g}_i \sim \mathcal{N}_{\mathbb{R}}(0, N^{-1}I)$ such that

$$\boldsymbol{v}_i = \frac{\widetilde{\boldsymbol{g}}_i}{\|\widetilde{\boldsymbol{g}}_i\|_2}\,,$$

Similarly to (4.6), we define

$$g_{ik} := \operatorname{sgn}(v_{ii})\, \widetilde{g}_{ik}\,, \qquad k \neq i\,,$$

and introduce an $N(0, N^{-1})$ variable g_{ii}, which is independent of the orthogonal matrix U and of $\widetilde{\boldsymbol{g}}_i$. Let $\boldsymbol{g}_i := (g_{i1}, \ldots, g_{iN})$ and note that $\boldsymbol{g}_i \sim \mathcal{N}_{\mathbb{R}}(0, N^{-1} I)$. Then we set $\boldsymbol{w}_i := \boldsymbol{e}_i + \boldsymbol{g}_i$ and $W_i := I - \boldsymbol{w}_i \boldsymbol{w}_i^*$ as before. With these modifications, we follow the proofs in Sects. 4–7 verbatim. The only difference is the derivation of (6.19). Instead of (6.12), we use the following integration by parts formula for real Gaussian random variables

$$\int_{\mathbb{R}} g f(g)\, \mathrm{e}^{-\frac{g^2}{2\sigma^2}}\, \mathrm{d}g = \sigma^2 \int_{\mathbb{R}} f'(g)\, \mathrm{e}^{-\frac{g^2}{2\sigma^2}}\, \mathrm{d}g\,, \tag{A.2}$$

for differentiable functions $f : \mathbb{R} \to \mathbb{R}$. Correspondingly, instead of (6.14), we have

$$\frac{\partial W_i}{\partial g_{ik}} = -\boldsymbol{e}_i \boldsymbol{e}_k^* - \boldsymbol{e}_k \boldsymbol{e}_i^* - \boldsymbol{e}_k \boldsymbol{g}_i^* - \boldsymbol{g}_i \boldsymbol{e}_k^*.$$

Hence, we get

$$\begin{aligned}
\frac{\partial \big(\widetilde{B}^{\langle i\rangle} G^{(i)}\big)_{kj}}{\partial g_{ik}} &= \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \big(\boldsymbol{e}_i \boldsymbol{e}_k^* + \boldsymbol{e}_k \boldsymbol{e}_i^* + \boldsymbol{e}_k \boldsymbol{g}_i^* + \boldsymbol{g}_i \boldsymbol{e}_k^*\big) \\
&\quad \times \widetilde{B}^{\langle i\rangle} \big(I - \boldsymbol{e}_i \boldsymbol{e}_i^* - \boldsymbol{e}_i \boldsymbol{g}_i^* - \boldsymbol{g}_i \boldsymbol{e}_i^* - \boldsymbol{g}_i \boldsymbol{g}_i^*\big) G^{(i)} \boldsymbol{e}_j \\
&\quad + \boldsymbol{e}_k^* \widetilde{B}^{\langle i\rangle} G^{(i)} \big(I - \boldsymbol{e}_i \boldsymbol{e}_i^* - \boldsymbol{e}_i \boldsymbol{g}_i^* - \boldsymbol{g}_i \boldsymbol{e}_i^* - \boldsymbol{g}_i \boldsymbol{g}_i^*\big) \\
&\quad \times \widetilde{B}^{\langle i\rangle} \big(\boldsymbol{e}_i \boldsymbol{e}_k^* + \boldsymbol{e}_k \boldsymbol{e}_i^* + \boldsymbol{e}_k \boldsymbol{g}_i^* + \boldsymbol{g}_i \boldsymbol{e}_k^*\big) G^{(i)} \boldsymbol{e}_j
\end{aligned}$$

instead of (6.15). Substitution into the identity

$$\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_j] = \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{g}_i}\big[g_{ik} (\widetilde{B}^{\langle i\rangle} G^{(i)})_{kj}\big] = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{g}_i}\left[\frac{\partial (\widetilde{B}^{\langle i\rangle} G^{(i)})_{kj}}{\partial g_{ik}}\right],$$

yields

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{g}_i}[\boldsymbol{g}_i^* \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{e}_j] &= \text{(r.h.s. of (8.15))} + \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\big(G^{(i)} (\widetilde{B}^{\langle i\rangle})^2 G^{(i)}\big)_{ji}\Big] \\
&\quad + \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\boldsymbol{e}_j^* G^{(i)} (\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{g}_i\Big] + \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\big(G^{(i)} \widetilde{B}^{\langle i\rangle} G^{(i)} \widetilde{B}^{\langle i\rangle}\big)_{ji}\Big] \\
&\quad + \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\boldsymbol{e}_j^* G^{(i)} \widetilde{B}^{\langle i\rangle} G^{(i)} \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i\Big] - \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\widehat{b}_i \big(G^{(i)} \widetilde{B}^{\langle i\rangle} G^{(i)}\big)_{ji}\Big] \\
&\quad - \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[\widehat{b}_j\, \boldsymbol{e}_j^* G^{(i)} \widetilde{B}^{\langle i\rangle} G^{(i)} \boldsymbol{g}_i\Big] \\
&\quad - \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[(G_{ij}^{(i)} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_j)\Big(\boldsymbol{e}_i^* (\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{g}_i + \boldsymbol{g}_i^* (\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{g}_i\Big)\Big] \\
&\quad - \frac{1}{N} \mathbb{E}_{\boldsymbol{g}_i}\Big[(G_{ij}^{(i)} + \boldsymbol{g}_i^* G^{(i)} \boldsymbol{e}_j)\Big(b_i^2 G_{ii}^{(i)} + \boldsymbol{g}_i^* (\widetilde{B}^{\langle i\rangle})^2 G^{(i)} \boldsymbol{e}_i\Big)\Big],
\end{aligned} \tag{A.3}$$

where we introduced $\widehat{b}_i := \boldsymbol{w}_i \widetilde{B}^{\langle i\rangle} \boldsymbol{w}_i$. Note that the last two terms were discussed in the unitary setup, and they were shown to be negligible. Therefore, to get (8.16) also in the orthogonal case, we rely on the following lemma to discard the supplementary small terms in (A.3). At first, let us discuss the case of $i = j$, which suffices for the proof of (2.20).

Lemma A.1. *Under the assumption of Proposition 6.1, we have the following bounds*

$$\begin{aligned}
&\big|\big(G^{(i)}(z) \widetilde{B}^{\langle i\rangle} G^{(i)}(z)\big)_{ii}\big| \prec \tfrac{1}{\eta}\,, && \big|\big(G^{(i)}(z) (\widetilde{B}^{\langle i\rangle})^2 G^{(i)}(z)\big)_{ii}\big| \prec \tfrac{1}{\eta}\,, \\
&\big|\boldsymbol{e}_i^* G^{(i)}(z) \widetilde{B}^{\langle i\rangle} G^{(i)}(z) \boldsymbol{g}_i\big| \prec \tfrac{1}{\eta}\,, && \big|\boldsymbol{e}_i^* G^{(i)}(z) (\widetilde{B}^{\langle i\rangle})^2 G^{(i)}(z) \boldsymbol{g}_i\big| \prec \tfrac{1}{\eta}\,, \\
&\big|\big(G^{(i)}(z) \widetilde{B}^{\langle i\rangle} G^{(i)}(z) \widetilde{B}^{\langle i\rangle}\big)_{ii}\big| \prec \tfrac{1}{\eta}\,, && \big|\boldsymbol{e}_i^* G^{(i)}(z) \widetilde{B}^{\langle i\rangle} G^{(i)}(z) \widetilde{B}^{\langle i\rangle} \boldsymbol{g}_i\big| \prec \tfrac{1}{\eta}\,,
\end{aligned} \tag{A.4}$$

for all $i \in [\![1, \ldots, N]\!]$.

Proof. We drop z from the notation. For the first two terms, we have

$$\left|\left(G^{(i)}(\widetilde{B}^{\langle i\rangle})^k G^{(i)}\right)_{ii}\right| \le \|\widetilde{B}^{\langle i\rangle}\|^k \|G^{(i)}\boldsymbol{e}_i\|_2^2 \lesssim \left((G^{(i)})^* G^{(i)}\right)_{ii} = \frac{\operatorname{Im} G^{(i)}_{ii}}{\eta} \prec \frac{1}{\eta}\,, \tag{A.5}$$

for $k = 1, 2$, where in the last step we used assumption (6.1). For the third and fourth terms, we have, for $k = 1, 2$,

$$\begin{aligned}\left|\boldsymbol{e}_i^* G^{(i)}(\widetilde{B}^{\langle i\rangle})^k G^{(i)}\boldsymbol{g}_i\right| &\le \|\widetilde{B}^{\langle i\rangle}\|^k \|G^{(i)}\boldsymbol{e}_i\|_2 \|G^{(i)}\boldsymbol{g}_i\|_2 \\ &\lesssim \left(\left((G^{(i)})^* G^{(i)}\right)_{ii}\right)^{\frac{1}{2}} \left(\boldsymbol{g}_i^*(G^{(i)})^* G^{(i)}\boldsymbol{g}_i\right)^{\frac{1}{2}} \\ &= \frac{1}{\eta}\left(\operatorname{Im} G^{(i)}_{ii}\right)^{\frac{1}{2}} \left(\operatorname{Im} \boldsymbol{g}_i^* G^{(i)}\boldsymbol{g}_i\right)^{\frac{1}{2}} \prec \frac{1}{\eta}\,,\end{aligned}$$

where in the last step we used assumption (6.1) and estimate (5.44). For the fifth term we note that $(G^{(i)}\widetilde{B}^{\langle i\rangle} G^{(i)}\widetilde{B}^{\langle i\rangle})_{ii} = b_i\left(G^{(i)}\widetilde{B}^{\langle i\rangle} G^{(i)}\right)_{ii}$ and the bound follows from (A.5). For the last term, we have

$$\left|\boldsymbol{e}_i^* G^{(i)}\widetilde{B}^{\langle i\rangle} G^{(i)}\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\right| \le \|\widetilde{B}^{\langle i\rangle}\| \|\boldsymbol{e}_i^* G^{(i)}\|_2 \|G^{(i)}\widetilde{B}^{\langle i\rangle}\boldsymbol{g}_i\|_2 \prec \frac{1}{\eta}\,, \tag{A.6}$$

where we used assumption (6.1) and estimate (5.44). This completes the proof.

All the other arguments in Sects. 4–7 work for the orthogonal case as well without modifications. This proves (2.20) of Theorem 2.5 for the Haar orthogonal case.
For (2.21), analogously to (A.4), we shall estimate the second to the seventh terms on the right side of (A.3), under the assumption of Proposition 8.3. To bound these terms, we can pursue the discussion from (A.5)–(A.6) with $\operatorname{Im} G^{(i)}_{ii}$ replaced by $\operatorname{Im} G^{(i)}_{jj}$ in the bounding procedure. Hence, it suffices to show for all $j \neq i$ that $|G^{(i)}_{jj}| \prec 1$. To see this, we use the fact $|G_{jj}| \prec 1$ from Theorem 2.5 and $|G_{jj} - G^{(i)}_{jj}| \prec \frac{1}{\sqrt{N\eta}}$ from Lemma 4.1 with $k = j$. Note that the assumptions of Lemma 4.1 are guaranteed by Theorem 2.5, (8.9) and assumption (8.3). Hence, (2.21) also holds in the orthogonal case.

Appendix B. Two Point Mass Case

In this section, we present our result when both, μ_α and μ_β, are convex combinations of two point masses. Without loss of generality (up to shifting and scaling), we may assume that μ_α and μ_β are of the following form,

$$\mu_\alpha = \xi\delta_1 + (1-\xi)\delta_0\,, \qquad \mu_\beta = \zeta\delta_\theta + (1-\zeta)\delta_0\,, \tag{B.1}$$

with real parameters ξ, ζ and θ satisfying

$$\theta \neq 0\,, \qquad \xi, \zeta \in \left(0, \frac{1}{2}\right], \qquad \xi \le \zeta\,, \qquad (\theta, \xi, \zeta) \neq \left(-1, \frac{1}{2}, \frac{1}{2}\right).$$

Here we excluded the case $(\theta, \xi, \zeta) = (-1, \frac{1}{2}, \frac{1}{2})$ since it is equivalent to $(\theta, \xi, \zeta) = (1, \frac{1}{2}, \frac{1}{2})$ under a shifting, where the latter is a special case of $\mu_\alpha = \mu_\beta$. In Section 7 of [1], we explained why the setting of (B.1) is special, and we thus excluded it from Theorem 2.5.
Following [24], we argued in Lemma 7.1 of [1] that in the setting of (B.1) we have

$$\mathcal{B}_{\mu_\alpha \boxplus \mu_\beta} = (\ell_1, \ell_2) \cup (\ell_3, \ell_4)\,, \tag{B.2}$$

in case $\mu_\alpha \neq \mu_\beta$, while we have

$$\mathcal{B}_{\mu_\alpha \boxplus \mu_\alpha} = (\ell_1, \ell_4)\,, \tag{B.3}$$

in case $\mu_\alpha = \mu_\beta$, where

$$\ell_1 := \min\left\{\frac{1}{2}\left(1 + \theta - \sqrt{(1-\theta)^2 + 4\theta r_+}\right), \frac{1}{2}\left(1 + \theta - \sqrt{(1-\theta)^2 + 4\theta r_-}\right)\right\},$$

$$\ell_2 := \max\left\{\frac{1}{2}\Big(1+\theta-\sqrt{(1-\theta)^2+4\theta r_+}\Big), \frac{1}{2}\Big(1+\theta-\sqrt{(1-\theta)^2+4\theta r_-}\Big)\right\},$$
$$\ell_3 := \min\left\{\frac{1}{2}\Big(1+\theta+\sqrt{(1-\theta)^2+4\theta r_+}\Big), \frac{1}{2}\Big(1+\theta+\sqrt{(1-\theta)^2+4\theta r_-}\Big)\right\},$$
$$\ell_4 := \max\left\{\frac{1}{2}\Big(1+\theta+\sqrt{(1-\theta)^2+4\theta r_+}\Big), \frac{1}{2}\Big(1+\theta+\sqrt{(1-\theta)^2+4\theta r_-}\Big)\right\},$$

with $r_\pm := \xi+\zeta-2\xi\zeta \pm \sqrt{4\xi\zeta(1-\xi)(1-\zeta)}$.
In Remark 7.2 of [1] we argued that, in the case $\mu_\alpha = \mu_\beta$, the point $E = 1 \in \mathcal{B}_{\mu_\alpha \boxplus \mu_\alpha}$ is special in the sense that $m_{\mu_\alpha\boxplus\mu_\alpha}(1+\mathrm{i}0)$ is unstable under small perturbations. We thus expect a modified local law in neighborhoods of this special point. To proceed we need some more notation. Recall the domains $\mathcal{S}_{\mathcal{I}}(a,b)$ in (2.18). For given (small) $\varsigma, \gamma > 0$, we set

$$\mathcal{S}^{\varsigma}_{\mathcal{I}}(a,b) := \left\{z \in \mathcal{S}_{\mathcal{I}}(a,b) : \varsigma|z-1| \geq \max\left\{\sqrt{\mathrm{d_L}(\mu_A,\mu_\alpha)}, \sqrt{\mathrm{d_L}(\mu_B,\mu_\beta)}\right\}\right\}$$

and

$$\widetilde{\mathcal{S}}^{\varsigma}_{\mathcal{I}}(a,b) := \mathcal{S}^{\varsigma}_{\mathcal{I}}(a,b) \cap \left\{z \in \mathbb{C} : |z-1| \geq \frac{N^\gamma}{(N\eta)^{\frac{1}{4}}}\right\}. \tag{B.4}$$

The following proposition presents the local law under the setting (B.1).

Proposition B.1. *Let μ_α, μ_β be as in (B.1), with fixed ξ, ζ and θ. Assume that the empirical eigenvalue distributions μ_A, μ_B of the sequences of matrices A, B satisfy (2.14). Fix any compact nonempty interval $\mathcal{I} \subset \mathcal{B}_{\mu_\alpha\boxplus\mu_\beta}$. With the notations and assumptions of Theorem 2.5, we have the following conclusions:*

(i) *If $\mu_\alpha \neq \mu_\beta$, then, for any fixed $\gamma > 0$,*

$$\max_{1\leq i\leq N}\left|G_{ii}(z) - (a_i - \omega_B(z))^{-1}\right| \prec \frac{1}{\sqrt{N\eta}},$$
$$|\omega^c_A(z) - \omega_A(z)| \prec \frac{1}{\sqrt{N\eta}}, \qquad |\omega^c_B(z) - \omega_B(z)| \prec \frac{1}{\sqrt{N\eta}},$$

hold uniformly for all $z \in \mathcal{S}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$.

(ii) *If $\mu_\alpha = \mu_\beta$, then, for sufficiently small $\varsigma > 0$ and any fixed $\gamma > 0$,*

$$\max_{1\leq i\leq N}\left|G_{ii}(z) - (a_i - \omega_B(z))^{-1}\right| \prec \frac{1}{|z-1|}\frac{1}{\sqrt{N\eta}}$$
$$|\omega^c_A(z) - \omega_A(z)| \prec \frac{1}{|z-1|}\frac{1}{\sqrt{N\eta}}, \qquad |\omega^c_B(z) - \omega_B(z)| \prec \frac{1}{|z-1|}\frac{1}{\sqrt{N\eta}}, \tag{B.5}$$

hold uniformly for all $z \in \widetilde{\mathcal{S}}^{\varsigma}_{\mathcal{I}}(\eta_{\mathrm{m}}, 1)$.

Proof. Recall the notation $\Gamma_{\mu_1,\mu_2}(\omega_1,\omega_2)$ in (3.8). In [1] (see the proof of Proposition 7.4 therein), we proved that under the setting (B.1) and assumption (2.14), one has the following results on the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A,\omega_B,z) = 0$: There exists a positive constant S such that the following two estimates hold.

(i) If $\mu_\alpha \neq \mu_\beta$, we have

$$\Gamma_{\mu_A,\mu_B}(\omega_A,\omega_B) \leq S, \qquad z \in \mathcal{S}_{\mathcal{I}}(0,1),$$

and (3.9), (3.14) and (3.15) hold on $\mathcal{S}_{\mathcal{I}}(0,1)$.

(ii) If $\mu_\alpha = \mu_\beta$, we have

$$\Gamma_{\mu_A,\mu_B}(\omega_A,\omega_B) \leq \frac{S}{|z-1|}, \qquad z \in \mathcal{S}^{\varsigma}_{\mathcal{I}}(0,1), \tag{B.6}$$

and (3.14) and (3.15) hold on $\mathcal{S}^{\varsigma}_{\mathcal{I}}(0,1)$, while (3.9) holds on $\mathcal{S}^{\varsigma}_{\mathcal{I}}(0,1)$ with S replaced by $\frac{S}{|z-1|}$.

Note that, the proofs of Lemma 5.1, Lemma 6.1 and Lemma 7.1 still work since we have the bounds (3.9), (3.14), and (3.15) as well. Although the bound in (3.9) should be replaced by $\frac{2S}{|z-1|}$ in the case $\mu_\alpha = \mu_\beta$, it is harmless for our proof. Hence, analogously to the proof of Theorem 2.5, one can use Lemma 3.3, Lemma 7.1 and estimates (3.9), (3.14) and (3.15), to complete the proof of Proposition B.1. Especially, the proof in the case $\mu_\alpha \neq \mu_\beta$ exactly agrees with the proof of Theorem 2.5.

For the case $\mu_\alpha = \mu_\beta$, we need to replace S by $\frac{S}{|z-1|}$ in Lemma 3.3 due to (B.6). In the sequel, we simply illustrate the continuity argument in this case. Let $z, z' \in \widetilde{\mathcal{S}}^S_{\mathcal{I}}(a,b)$, where $z = E + \mathrm{i}\eta$ and $z' = E + \mathrm{i}\eta'$, with $\eta' = \eta + N^{-5}$. In addition, we set $z_0 = z$, $\omega_1 = \omega_A$, $\omega_2 = \omega_B$, $\widetilde{\omega}_1 = \omega_A^c$ and $\widetilde{\omega}_2 = \omega_B^c$ in Lemma 3.3. Suppose now that (B.5) holds for z'. Using the Lipschitz continuity of the Green function (*i.e.* $\|G(z) - G(z')\| \leq N^2|z - z'|$) and of the subordination functions $\omega_A(z)$ and $\omega_B(z)$ (*c.f.* (3.9) with S replaced by $\frac{S}{|z-1|}$), we can choose δ in (3.11) to be

$$\delta = \frac{N^\gamma}{|z-1|}\frac{1}{\sqrt{N\eta}} + O(N^{-3}), \tag{B.7}$$

In light of the condition $k^2 > \delta K \frac{S}{|z-1|}$ (*c.f.* sentence above (3.12), with S replaced by $\frac{S}{|z-1|}$), one needs to guarantee that $\delta S \leq |z-1|\varepsilon$, for sufficiently small constant $\varepsilon > 0$, which is a direct consequence of the assumption that $z \in \widetilde{\mathcal{S}}^S_{\mathcal{I}}(a,b)$ and (B.7). Note that $\|\widetilde{r}(z)\|_2 \prec \frac{1}{\sqrt{N\eta}}$ remains valid since estimate (7.15) does not depend on the stability of the system $\Phi_{\mu_A,\mu_B}(\omega_A, \omega_B, z) = 0$, as long as (3.14), (3.15) and (3.9) hold. The remaining parts of the proof are analogous to those of Theorem 2.5 and we thus omit the details.

References

1. Bao, Z.G., Erdős, L., Schnelli, K.: Local stability of free additive convolution. J. Funct. Anal. **271**(3), 672–719 (2016)
2. Bao, Z.G., Erdős, L., Schnelli, K.: On the local single ring theorem **(in preparation)**
3. Belinschi, S.: A note on regularity for free convolutions. Ann. Inst. Henri Poincaré Probab. Stat. **42**(5), 635–648 (2006)
4. Belinschi, S.: The Lebesgue decomposition of the free additive convolution of two probability distributions. Probab. Theory Related Fields **142**(1–2), 125–150 (2008)
5. Belinschi, S.: L^∞-boundedness of density for free additive convolutions. Rev. Roumaine Math. Pures Appl. **59**(2), 173–184 (2014)
6. Belinschi, S., Bercovici, H.: A new approach to subordination results in free probability. J. Anal. Math. **101**(1), 357–365 (2007)
7. Belinschi, S., Bercovici, H., Capitaine, M., Février, M.: Outliers in the spectrum of large deformed unitarily invariant models (2014). arXiv:1412.4916
8. Benaych-Georges, F.: Local single ring theorem (2015). arXiv:1501.07840
9. Bercovici, H., Voiculescu, D.: Free convolution of measures with unbounded support. Indiana Univ. Math. J. **42**, 733–773 (1993)
10. Bercovici, H., Voiculescu, D.: Regularity questions for free convolution, nonselfadjoint operator algebras, operator theory, and related topics. Oper. Theory Adv. Appl. **104**, 37–47 (1998)
11. Biane, P.: Representations of symmetric groups and free probability. Adv. Math. **138**(1), 126–181 (1998)
12. Bourgade, P., Erdős, L., Yau, H.-T., Yin, J.: Fixed energy universality for generalized Wigner matrices. Commun. Pure Appl. Math. **69**(10), 1815–1881 (2016)
13. Capitaine, M.: Additive/multiplicative free subordination property and limiting eigenvectors of spiked additive deformations of Wigner matrices and spiked sample covariance matrices. J. Theor. Probab. **26**(3), 595–648 (2013)
14. Chatterjee, S.: Concentration of Haar measures, with an application to random matrices. J. Funct. Anal. **245**(2), 379–389 (2007)
15. Chistyakov, G.P., Götze, F.: The arithmetic of distributions in free probability theory. Central Eur. J. Math. **9**, 997–1050 (2011)
16. Collins, B.: Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability. Int. Math. Res. Notices **2003**(17), 953–982 (2003)
17. Diaconis, P., Shahshahani, M.: The subgroup algorithm for generating uniform random variables. Probab. Eng. Inform. Sci. **1**(01), 15–32 (1987)
18. Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. Ann. Henri Poincaré **14**, 1837–1926 (2013)

19. Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. Probab. Theory Related Fields **154**(1–2), 341–407 (2012)
20. Erdős, L., Schnelli, K.: Universality for random matrix flows with time-dependent density (2015). arXiv:1504.00650
21. Erdős, L., Yau, H.-T.: Universality of local spectral statistics of random matrices. Bull. Am. Math. Soc. **49**(3), 377–414 (2012)
22. Guionnet, A., Krishnapur, M., Zeitouni, O.: The single ring theorem. Ann. Math. (2) **174**, 1189–1217 (2011)
23. Hiai, F., Petz, D.: The semicircle law, free random variables and entropy. Math. Surveys Monogr. **77**. Amer. Math. Soc., Providence (2000)
24. Kargin, V.: On eigenvalues of the sum of two random projections. J. Stat. Phys. **149**(2), 246–258 (2012)
25. Kargin, V.: A concentration inequality and a local law for the sum of two random matrices. Prob. Theory Related Fields **154**, 677–702 (2012)
26. Kargin, V.: Subordination for the sum of two random matrices. Ann. Probab. **43**(4), 2119–2150 (2015)
27. Landon, B., Yau, H.-T.: Convergence of local statistics of Dyson Brownian motion (2015). arXiv:1504.03605
28. Mezzadri, F.: How to generate random matrices from the classical compact groups. Notices Am. Math. Soc. **54**(5), 592–604 (2007)
29. Pastur, L., Vasilchuk, V.: On the law of addition of random matrices. Commun. Math. Phys. **214**(2), 249–286 (2000)
30. Speicher, R.: Free convolution and the random sum of matrices. Publ. Res. Inst. Math. Sci. **29**(5), 731–744 (1993)
31. Voiculescu, D.: Limit laws for random matrices and free products. Invent. Math. **104**(1), 201–220 (1991)
32. Voiculescu, D., Dykema, K.J., Nica, A.: Free random variables. CRM Monogr. Ser. Amer. Math. Soc., Providence (1992)

Communicated by H.-T. Yau